The strategy of conflict
Prospectus for a reorientation of game theory

THOMAS C. SCHELLING
*Harvard University**

On the strategy of pure conflict—the zero-sum games—*game theory* has yielded important insight and advice. But on the strategy of action where conflict is mixed with mutual dependence—the non-zero-sum games involved in wars and threats of war, strikes, negotiations, criminal deterrence, class war, race war, price war, and blackmail; maneuvering in a bureaucracy or a social hierarchy or in a traffic jam; and the coercion of one's own children—traditional game theory has not yielded comparable insight or advice. These are the "games" in which, though the element of conflict provides the dramatic interest, mutual dependence is part of the logical structure and demands some kind of collaboration or mutual accommodation—tacit, if not explicit—even if only in the avoidance of mutual disaster. These are also games in which, though secrecy may play a strategic role, there is some essential need for the signaling of intentions and the meeting of minds. Finally, they are games in which what one player *can* do to avert mutual damage affects what another player *will* do to avert it, so that it is not always an advantage to possess initiative, knowledge, or freedom of choice.

Traditional game theory has, for the most part, applied to these mutual-dependence games (non-zero-sum games) the methods and concepts that proved successful in studying the strategy of pure conflict. The present paper attempts to enlarge the scope of game theory, taking the zero-sum game to be a limiting case rather than a point of departure. The proposed extension of the theory will be mainly along two lines. One is to identify the perceptual and suggestive element in the formation of mutually consistent expectations. The other is to identify some of the basic "moves" that may occur in actual games of strategy and the structural elements that the moves depend on; it involves such concepts as "threat," "enforcement," and the capacity to communicate or to destroy communication.

That game theory is underdeveloped along these two lines may reflect its preoccupation with the zero-sum game. Suggestions and inferences, threats and promises, are of no consequence in the accepted theory of zero-sum games. They are of no consequence because they imply a relationship between the two players that, unless perfectly innocuous, must be to the disadvantage of one player; and he can destroy it by adopting a minimax strategy, based, if necessary, on a randomizing mechanism. So the "rational strategies" pursued by two players in a situation of pure conflict—as typified by pursuit and evasion—should not be expected to reveal what kind of behavior is conducive to mutual accommodation or how mutual dependence can be exploited for unilateral gain.

* The author is currently with The RAND Corporation, on leave from Harvard University.

I. *Mutual Perception and Suggestive Behavior*

If the zero-sum game is the limiting case of pure conflict, what is the other extreme? It must be the "pure-collaboration" game in which the players win or lose together, having identical preferences regarding the outcome. Whether they win fixed shares of the total or shares that vary with the joint total, they must rank all possible outcomes identically in their separate preference scales. (And, to avoid any initial conflict, it has to be evident to the players that the preferences are identical, so that there is no conflict of interest in the information or misinformation that they try to convey to each other.)

What is there about pure collaboration that relates it to game theory or to bargaining? A partial answer, just to establish that this game is not trivial, is that it may contain problems of perception and communication of a kind that quite generally occurs in non-zero-sum games. Whenever the communication structure does not permit players to divide the task ahead of time according to an explicit plan, it may not be easy to co-ordinate behavior in the course of the game. Players have to understand each other, to discover patterns of individual behavior that make each player's actions predictable to the other; they have to test each other for a shared sense of pattern or regularity and to exploit clichés, conventions, and impromptu codes for signaling their intentions and responding to each other's signals. They must communicate by hint and by suggestive behavior. Two vehicles trying to avoid collision, two people dancing together to unfamiliar music, or members of a guerrilla force that become separated in combat have to concert their intentions in this fashion, as do the applauding members of a concert audience, who must at some point "agree" on whether to press for an encore or taper off together.

If *chess* is the standard example of a zero-sum game, *charades* may typify the game of pure co-ordination; if *pursuit* epitomizes the zero-sum game, *rendezvous* may do the same for the co-ordination game.

An experiment of O. K. Moore and M. I. Berkowitz provides a nice mixture in which the two limiting cases are both visible. It involves a zero-sum game between two teams, each team consisting of three people. The three members of the team have identical interests but, because of a special feature of the game, cannot behave as a single entity. The special feature is that the three members of each team are separated and can communicate only by telephone and that all six telephones are connected on the same line so that everyone can hear both the other team and his own teammates. No prearrangement of codes is permitted. Between teams we have here a pure-conflict game; among the members of the team we have a pure-co-ordination game (25).

If in this game we suppress the "other team" and if the three players simply try to co-ordinate a winning strategy in a game of skill or chance in the face of communication difficulty, we have a three-person pure-co-ordination game. Several "games" of this sort have been studied, both experimentally and formally; in fact, there is substantial overlap at this point between the non-zero-sum game and organization or communication theory.[1]

The author's experiments, reported in an earlier issue of this *Journal*, have shown that

[1] An extensive formal analysis of the co-ordination problem is developed by Marschak (22, 24) and with Roy Radner (23). Examples of interesting and relevant empirical work can be found in Bavelas (2), Heise and Miller (14), Mueller (20), and Carmichael, Hogan, and Walter (4).

co-ordinated choice is possible even in the complete absence of communication. Further, they showed that there are tacit bargaining situations in which the *conflict* of interest in the choice of action may be overwhelmed by the sheer need for concerting on *some* action; in those situations, the limiting case of pure co-ordination isolates the essential feature of the corresponding non-zero-sum game (31).

So we do have, in this *co-ordinated problem-solving*, with its dependence on the conveyance and perception of intentions or plans, a phenomenon that brings out an essential aspect of the non-zero-sum game; and it stands in much the same relation to it as the zero-sum game, namely, that of "limiting case." One is the mixed conflict-co-operation game with all scope for co-operation eliminated; the other is the mixed conflict-co-operation game with the conflict eliminated. In one the premium is on secrecy, in the other on revelation.

It is to be stressed that the pure-co-ordination game is a *game of strategy* in the strict technical sense. It is a behavior situation in which each player's best choice of action depends on the action he expects the other to take, which he knows depends, in turn, on the other's expectations of his own. This interdependence of expectations is precisely what distinguishes a game of strategy from a game of chance or a game of skill. In the pure-co-ordination game the interests are convergent; in the pure-conflict game the interests are divergent; but in neither case can a choice of action be made wisely without regard to the dependence of the outcome on the mutual expectations of the players.[2]

Recall the famous case of Holmes and Moriarty on separate trains, neither directly in touch with the other, each having to choose whether to get off at the next station. We can consider three kinds of payoff. In one, Holmes wins a prize if they get off at different stations, Moriarty wins it if they get off at the same station; this is the zero-sum game, in which the preferences of the two players are perfectly correlated inversely. In the second case, Holmes and Moriarty will both be rewarded if they succeed in getting off at the same station, whatever station that may be; this is the pure-co-ordination game, in which the preferences of the players are perfectly correlated positively. The third payoff would show Holmes and Moriarty both being rewarded if they succeed in getting off at the same station, but Holmes gaining more if both he and Moriarty get off at one particular station, Moriarty gaining more if both get off at some other particular station, both losing unless they get off at the same station. This is the usual non-zero-sum game, or "imperfect-correlation-of-preferences" game. This is the mixture of conflict and mutual dependence that epitomizes bargaining situations. By specifying particular communication and intelligence systems for the players, we can enrich the game or make it trivial or pro-

[2] Concerning this point, Kaysen (16) in his review of the Von Neumann and Morgenstern book says: "The theory of such games of strategy deals precisely with the actions of several agents, in a situation in which all actions are interdependent, and where, in general, there is no possibility of what we called parametrization that would enable each agent (player) to behave as if the actions of the others were given. In fact, it is this very lack of parametrization which is the essence of a game." Similar language is used by Luce and Raiffa (21, p. 14): "Intuitively, the problem of conflict of interest is, for each participant, a problem of individual decision making under a mixture of risk and uncertainty, the uncertainty arising from his ignorance as to what the others will do." Their preoccupation is with the conflict, however; the case of coincident preferences among the players they dispose of as trivial (pp. 59, 88), and they deal with such players as a single individual (p. 13).

vide an advantage to one of the two players in the first and third variants.

The essential game-of-strategy element is present in all three cases: the best choice for either depends on what he expects the other to do, knowing that the other is similarly guided, so that each is aware that each must try to guess what the second guesses the first will guess the second to guess and so on, in the familiar spiral of reciprocal expectations.

A RECLASSIFICATION OF GAMES

Before going further, we can usefully reclassify game situations. The twofold division into zero-sum and non-zero-sum lacks the symmetry that we need and fails to identify the limiting case that stands opposite to the zero-sum game. The essentials of a classification scheme for a two-person game could be represented on a two-dimensional diagram. The values of any particular outcome of the game, for the two players, would be represented by the two co-ordinates of a point. All possible outcomes of a pure-conflict game would be represented by some or all of the points on a negatively inclined line, those of a pure common-interest game by some or all of the points on a positively inclined line. In the mixed game, or bargaining situation, at least one pair of points would denote a negative slope and at least one pair a positive slope.[3]

We could stay close to traditional terminology, with respect to the strictly pure games, by calling them *fixed-sum* and *fixed-proportions* games, getting the unwieldy *variable-sum–variable-proportions* as the name for all games except the limiting cases. We could also call them perfect-negative-correlation games and perfect-positive-correlation games, referring to the correlation of their preferences with respect to outcomes, leaving for the richer mixed game the rather dull title of "imperfect-correlation game."

---

[3] If the nature of the game makes it desirable for a player to use a random device in the choice of his strategy, or feasible for the players to negotiate an enforcible agreement that, like a drawing of lots, depends on a chance mechanism, there may be room for co-operation in the choice of *strategies* even when there is perfect disagreement over the ranking of *outcomes*. In that case the points representing the pure-conflict game must meet the tighter restriction of lying on a straight line, with the two axes measuring the players' "utilities" in the sense now familiar in game theory. This restriction also applies to the pure common-interest game, since players who agree perfectly on the ranking of *outcomes* may not agree on the desirability of, say, one particular point over a fifty-fifty chance between the two points immediately above and below it. Thus "strictly pure" conflict and common-interest games, providing no scope for collaboration in the one case and no scope for disagreement in the other, would have to show the *expected values* of all pertinent mixed (random) strategies lying along the downward-sloping and upward-sloping lines, respectively, with axes measured in "utility units" of the kind mentioned; this in turn means that the points denoting *outcomes* must lie on a *straight* line.

Also, the pure games cannot admit "side payments." If one of the partners in a pure common-interest game threatens to sabotage the effort unless he is paid—assuming that the communication and enforcement structure of the game makes this possible—a conflict of interest is introduced; in effect, the point denoting the payment of a bribe would appear to the upper left or lower right of another point or points on the upward-sloping line, producing the configuration of a mixed game. And if one of the players in a pure-conflict game can threaten damage or offer compensation to induce his opponent to yield in this game, there is scope for bargaining; there is no longer a relationship of pure conflict, and the points denoting the threatened damage or promised compensation would lie off the downward-sloping line. In other words, *all* pertinent potential outcomes must be allowed for. (Two simultaneous pure-conflict games, even if they meet the restriction of straight lines, provide room for negotiation unless the slopes of the two lines happen to be identical.)

The difficulty is in finding a sufficiently rich name for the mixed game in which there is both conflict and mutual dependence. It is interesting that we have no very good word for the *relationship* between the players: in the common-interest game we can refer to them as "partners" and in the pure-conflict game as "opponents" or "adversaries"; but the mixed relationship that is involved in wars, strikes, negotiations, and so forth, requires a more ambivalent term.[4] In the rest of this paper I shall refer to the mixed game as a *bargaining game* or *mixed-motive game,* since these terms seem to catch the spirit. "Mixed-motive" refers not, of course, to an individual's lack of clarity about his own preferences but rather to the ambivalence of his relationship to the other player—the mixture of mutual dependence and conflict, of partnership and competition. "Non-zero-sum" refers to the mixed game together with the pure common-interest game. And because it characterizes the problem and the activity involved, *co-ordination game* seems a good name for the perfect sharing of interests.

GAMES OF CO-ORDINATION

While most of this paper will be about the mixed game, a brief discussion of the pure co-ordination game is in order, partly to indicate that it is an important game in its own right, partly to identify certain qualities of the mixed game that appear most clearly in this limiting case.

An illustration of the game—and the fact that people can play it with success—is the following experiment recorded by the author (31, pp. 21–22) in an earlier issue of this *Journal.* Forty-two people were asked to choose heads or tails without any communication among themselves. Each participant was instructed to try to concert with the rest, success being represented by the proportion of the total sample giving the same answer that he did. Evidently on the basis of chance a member of the 42-person sample might have expected to concert his choice successfully with about 20 or 21 other participants.

Thirty-six chose heads. The probability of as many as 36 heads or tails in a sample of 42, on the hypothesis that people could do no better than chance, is less than .0001. In similar fashion respondents were able without communication to concert their choices of points on a map, candidates on a ballot, numbers in a series of numbers, places to meet in a specified city, or a division of a sum into two piles. When instructed, "Write some positive number—if you all write the same number, you win," out of the infinity of possible numbers to choose, 40 per cent were able to pick the same number. Each of the problems evidently provided some focal point for a concerted choice, some clue to co-ordination, some rationale for the convergence of the participants' mutual expectations.

In the earlier article it was argued that the same kind of co-ordinating clue—usually a "non-mathematical" clue—might be a potent force not only in pure co-ordination but in the mixed situation that includes conflict; and, in fact, the experiments demon-

[4] It deserves to be emphasized that non-zero-sum games can as properly be classed under theory of partnership as under theory of conflict; and, for providing insight into problems like that of limiting war, there is merit in using words that bring out the common interest of the adversaries and the "bargaining process" involved in the military maneuvers themselves. As the author has pointed out in another paper (34), even the problem of surprise attack is logically equivalent to a problem in partnership discipline. If *theory of games* has become endowed with a too conflict-oriented connotation, perhaps something like *theory of interdependent decision* would be a neutral term that equally covers the two limiting cases as well as the mixed case.

strated that, in the complete absence of communication, this is certainly true. But there are a number of instances in which pure co-ordination itself—the *tacit* procedure of identifying partners and concerting plans with them—is a significant phenomenon. A good example is the formation of riotous mobs.

It is usually the essence of mob formation that the potential members have to know not only where and when to meet but just when to act so that they act in concert. Overt leadership solves the problem; but leadership can often be identified and eliminated by the authority trying to prevent mob action. In this case the mob's problem is to act in unison without overt leadership, to find some common signal that makes everyone confident that if he acts on it, he will not be acting alone. The role of "incidents" can thus be seen as a co-ordinating role; it is a substitute for overt leadership and communication. Without something like an incident, it may be difficult to get action at all, since immunity requires that all know when to act together. Similarly, the city that provides no "obvious" central point or dramatic site may be one in which mobs find it difficult to congregate spontaneously; there is no place so "obvious" that it is evident to everyone that it is obvious to everyone else. Bandwagon behavior, in the selection of leadership or in voting behavior, may also depend on "mutually perceived" signals, when a part of each person's preference is a desire to be in a majority or, at least, to see some majority coalesce.[5]

Excessively polarized behavior may be the unhappy result of dependence on tacit co-ordination and maneuver. When whites and Negroes see that an area will "inevitably" become occupied exclusively by Negroes, the "inevitability" is a feature of convergent expectation.[6] What is most directly perceived as inevitable is not the final result but the *expectation* of it, which, in turn, makes the result inevitable. Everyone expects everyone else to expect everyone else to expect the result; and everyone is powerless to deny it. There is no stable focal point except at the extremes. Nobody can expect the tacit process to stop at 10, 30, or 60 per cent; no *particular* percentage commands agreement or provides a rallying point. If tradition suggests 100 per cent, tradition could be contradicted only by explicit agreement; if co-ordination has to be tacit, compromise may be impossible. People are at the mercy of a faulty communication system that makes it easy to "agree" (tacitly) to move but impossible to agree to stay. Quota systems in housing developments, schools, etc., can be viewed as efforts to substitute an explicit game with communication and enforcement for a tacit game that has an undesirably extreme "solution."

The co-ordination game probably lies behind the stability of institutions and traditions and perhaps the phenomenon of leadership itself. Among the possible sets of rules that might govern a conflict, tradition points to the particular set that everyone can expect everyone else to be conscious of as a conspicuous candidate for adoption; it wins by default over those that cannot readily be identified by tacit consent. The force of many rules of etiquette and social re-

[5] A closely related phenomenon is appreciated by the person who tries to blend into the crowd to avoid being called on to recite, picked on by a bully, or singled out for "election" to some post that everybody wants to escape.

[6] The phenomenon, called "tipping," is analyzed by Grodzins (12). A more innocuous example of explosively convergent expectations, based on tacit communication that has an almost electric quality, is the snicker that ignites an outburst of uncontrollable laughter in a nervous crowd.

straint, including some (like the rule against ending a sentence with a preposition) that have been divested of their relevance or authority, seems to depend on their having become "solutions" to a co-ordination game: everyone expects everyone to expect everyone to expect observance, so that non-observance carries the pain of conspicuousness. Clothing styles and motorcar fads may also reflect a game in which people do not wish to be left out of any majority that forms and are not organized to keep majorities from forming. The concept of *role* in sociology, which explicitly involves the expectations that others have about one's behavior, as well as one's expectations about how others will behave toward him, can in part be interpreted in terms of the stability of "convergent expectations," of the same type that is involved in the co-ordination game. One is trapped in a particular role, or by another's role, because it is the only role that in the circumstances can be identified by a process of tacit consent.

A good example might be the *esprit de corps* (or lack of it) of an army unit or naval vessel or the value system of a particular college or fraternity. These are social organisms that are subject to a substantial rate of replacement but that maintain their own peculiar identities to an extent that does not seem to be accounted for by selective or biased recruitment. The individual character of one of these units seems to be largely a matter of convergent expectations—everyone's expectation of what everyone expects of everyone—with the new arrivals' expectations being molded in time to help mold the expectations of subsequent arrivals. There is a sense of "social contract," the particular terms of which are sensed and accepted by each incoming generation. I am told that this persistence of a tradition in a social entity is one of the reasons why the legal identity of an army division or regiment—its name and number and history—is often deliberately preserved when its strength has fallen to where abolition might seem indicated: the tradition that goes with the legal identity of the group is an asset worth preserving for a future buildup. It may be the same phenomenon that makes it possible to collect income tax in some countries and not in others: if appropriate mutual expectations exist, people will expect evasion to be on a scale small enough not to overwhelm the authorities and may consequently pay up either out of a sense of reciprocated honesty or out of fear of apprehension, thus together justifying their own expectations.

Nature of the intellectual process in co-ordination. It should be emphasized that co-ordination is not a matter of guessing what the "average man" will do. One is not, in tacit co-ordination, trying to guess what another will do in an objective situation; one is trying to guess what the other will guess one's self to guess the other to guess, and so on ad infinitum. ("Meeting" someone in the personal column of a newspaper is a good example.) The reasoning becomes disconnected from the objective situation, except insofar as the objective situation may provide some clue for a concerted choice. The analogy is not just trying to vote with the majority but trying to vote with a majority when everyone wants to be in a majority and everyone knows it—not to predict Miss Rheingold of 1958 but to buy the stock or real estate that everyone expects everyone to expect everyone to buy. Investment in diamonds may be a perfect example; the greatest example of all may be the monetary role of gold, which can perhaps be explained only as the "solution" of a co-ordination game. (A common household version of the co-ordination game occurs when two people are cut off in a telephone conversation; if they both call back, they only get busy signals.)

Consider the game of "name a positive number." Experiments demonstrate that most people, asked just to pick a number, will pick numbers like 3, 7, 13, 100, and 1. But when asked to pick the same number that others will pick when the others are equally interested in picking the same number, and everyone knows that everyone else is trying, the motivation is different. The preponderant choice is the number 1. And there seems to be good logic in this: there is no unique "favored number"; the variety of candidates like 3, 7, etc., is embarrassingly large, and there is no good way of picking the "most favorite" or most conspicuous.

1	0	0	0	0
0	1	0	0	0
0	0	1	0	0
0	0	0	1	0
0	0	0	0	1

FIG. 1

If one then asks what number, among all positive numbers, is most clearly unique, or what *rule of selection would lead to unambiguous results,* one may be struck with the fact that the universe of all positive numbers has a "first" or "smallest" number.[7]

Game-theory formulation of the co-ordination problem. The payoff matrix for a pure co-ordination problem would look something like that in Figure 1. One player chooses a row, the other a column; and they receive the rewards denoted by the numbers contained in the cell where their choices intersect. If to each choice of one player there corresponds a single choice for the other that "wins" for both of them, we can arrange columns so that all the winning cells lie along the diagonal. In those cells there are positive payoffs to both players, in the rest we can put zeros. (For our present purpose there is nothing lost by letting a single number stand in each cell for the payoff to both players.)

But we must rule out a possible axiom that might seem to be suggested by analogy with other game theories, namely, that (to use the term of Luce and Raiffa) the "labeling" of rows, columns, and players should make no difference to the outcome.[8]

---

[7] There is a widely quoted passage in Keynes (17, p. 156) that may be worth repeating in order to point out that, while it deals with exactly the problem dealt with here, its conception of the "solution" is *not* at all the same: "Professional investment may be likened to those newspaper competitions in which the competitors have to pick out the six prettiest faces from a hundred photographs, the prize being awarded to the competitor whose choice most nearly corresponds to the average preference of the competitors as a whole; so that each competitor has to pick, not those faces which he himself finds prettiest, but those which he thinks likeliest to catch the fancy of the other competitors, all of whom are looking at the problem from the same point of view. It is not a case of choosing those which, to the best of one's judgment, are really the prettiest, nor even those which average opinion genuinely thinks prettiest. We have reached the third degree where we devote our intelligence to anticipating what average opinion expects the average opinion to be. And there are some, I believe, who practice the fourth, fifth, and higher degrees." This class of games demonstrates, incidentally, that the usual correlation between parametric behavior and large numbers does not hold for tacit play. To adapt "parametrically" to the behavior of others requires that their behavior be observable, not conjectural; the non-parametric character of tacit co-ordination remains, no matter how large the number of players.

[8] Labeling of the *players* is explicitly ruled out by Luce and Raiffa (21, pp. 123–27) in discussing co-operative games, and in effect is ruled out by Nash (26, 27) in his symmetry assumption. Labeling of *strategies* for tacit or explicit non-zero-sum games is implicitly precluded by dealing only with games in normal form, i.e., the abstract version of them as represented by a payoff matrix (which is itself an *analytical* de-

It is precisely because strategies are "labeled" in some sense—i.e., have symbolic or connotative characteristics that transcend the mathematical structure of the game—that players can rise above sheer chance and "win" these games; and it is for that same reason that these games are interesting and important.

Even the game portrayed in Figure 1, which might seem to have a minimum of symbolic significance attached to rows and columns, is not a hard one to "win," i.e., for players to do substantially better on than chance would suggest. (If we give that same game an infinite series of rows and columns, it seems to become easier rather than harder. In that case it is formally identical with the game mentioned earlier, "Pick a positive number," but, because the "labeling" is different, there is less tendency for minorities to congregate at 3, 7, 13, etc.) Just forming the matrix prejudices the choice, since it focuses attention on "first," "middle," "last," and so forth.[9] If strategies are not given sequential labels, i.e., labels that can be ordered like numbers and alphabets, but are given individual names, and these are not presented in any particular order, it is the names that must co-ordinate choice.

And here it becomes emphatically clear that the intellectual processes of choosing a strategy in pure conflict and choosing a strategy of co-ordination are of wholly different sorts. At least this is so if one admits the "minimax" solution, randomized if necessary, in the zero-sum game. In the pure-co-ordination game, the player's objective is to make contact with the other player through some imaginative process of introspection, of searching for shared clues; in the minimax strategy of a zero-sum game—most strikingly so with randomized choice—one's whole objective is to avoid any meeting of minds, even an inadvertent one.[10]

---

vice, not part of the game, and hence provides no left-right, upper-lower, or numerical ordering of the actual strategies). A good example in which the labeling of *players* is the controlling factor is the interrupted telephone call mentioned earlier, with the problem of who should call back and who should wait for the call.

[9] This point is typical of a number of demonstrations in the author's experiments reported earlier, to the effect that the postulate regarding the "independence of irrelevant alternatives" cannot be credited in the tacit game and, for analogous reasons, should not be expected to hold in the explicit bargaining game. Potential outcomes can be relevant to the co-ordination of choice, though not themselves near to being chosen. For a statement and discussion of this postulate see Luce and Raiffa (21, p. 127).

[10] Randomized strategies may nevertheless be useful to achieve a co-ordinated *distribution* of votes, say, among a panel of candidates. If a 55 per cent majority exists and knows that it does, among a hundred voters; if two out of six candidates are congenial to it; and if the three candidates polling the largest numbers of votes become the board of directors, there is danger that unco-ordinated polling may concentrate too many votes on the first (or second) majority choice, leaving the minority two winning candidates with 22 votes apiece. But if each member of the majority flips a coin to cast his vote for one of his party's men, the likelihood of one's getting as few as 22 votes is only one chance in six. If the minority, too, lacks an overt means of collaborating and relies on a chance device, the majority's chances are excellent.

A partial randomized strategy may also be used to reduce an area of conflict. Suppose two people sit at North and East sides of a card table, are to move to another card table adjacent that is identically oriented, must choose without communication what seats they will take at the other table, and will win prizes of $1 apiece if they pick adjacent seats. This is an easy co-ordination problem; but let us subvert the incentives, by giving an additional $2 premium to the player who is on the other's right in the event they succeed in sitting next to each other. This game has no equilibrium point; interests do not converge; there is no seating arrangement that would not

To illustrate, suppose that I am to name one card in an ordinary deck of fifty-two and you are to guess which one I name. Traditional game theory gives guidance on how to make my choice on the assumption that I do not want you to outguess me; I can select at random and defy you to have a better than random chance of guessing what I name. But if the game is that I *do* want you to guess correctly and you know that I will try to pick one that facilitates your guess, the random device can only guarantee to make tacit co-operation impossible. Holmes can *destroy* the labeling of the stations by flipping a coin to decide where to get off the train; and Moriarty has only a fifty-fifty chance of guessing a coin. But in the common-interest version they must somehow *use* the labeling of the stations in order to do better than pure chance; and how to use it may depend more on imagination than on logic, more on poetry or humor than on mathematics. It is noteworthy that traditional game theory does not assign a "value" to this game: how well people can concert in this fashion is something that, though hopefully amenable to systematic analysis, cannot be discovered by reasoning a priori. This corner of game theory is *inherently* dependent on empirical evidence.[11]

It should particularly be noted that to assert the influence of "labels"—i.e., of the symbolic and connotative details of the game—and the dependence of the theory on empirical evidence does not involve the question of whether game theory is predictive or normative—concerned with generalizations about actual choice or the strategy of correct choice. The assertion here is *not* that people simply *are* affected by symbolic details but that they *should* be for the purpose of correct play. A normative theory must produce strategies that are at least as good as what people can do without them. More, it must not deny or expunge details of the game that can demonstrably benefit two or more players and that the players, consequently, should not expunge or ignore in their mutual interest. Two

---

give one an incentive to move. (Each may wish that he could promise to sit on the other's left, but cannot.) A random strategy yields each player a minimax value of $1. But if each decides where he would sit in the pure common-interest game, then flips a coin to see whether he does sit there or sits opposite, the players guarantee that they neither choose the same seat nor sit opposite each other and share equal chances of winning the premium. This is an equilibrium pair of (mixed) strategies, worth an expected value of $2 apiece.

---

[11] Incidentally, in cases like this we need only to consider the question of what *price* players would pay for a bit of co-ordinating information and what different information patterns yield what chances of co-ordinating to find ourselves in the middle of Marschak's *theory of teams* (22, 23, 24).

There is also a version of "prisoners' dilemma" for this game: two accomplices, apprehended before their alibi is prepared and interrogated separately, who must concert the alibis they invent or be revealed in their guilt. A tantalizing variant can be built by supposing that confession carries a lighter sentence than unconfessed guilt; each player has a "minimax" strategy of confession and must not only consider which particular alibi constitutes the *best* alibi strategy but *how good it is* (in terms of likely coincidence with his partner's) and whether they share the decision to try it. The matrix might be:

.5 .5	0 .5	0 .5	0 .5
.5 0	1	0	0
.5 0	0	1	0
.5 0	0	0	1

(Lower left entry in each cell is payoff to player choosing row, upper right to player choosing column.)

couples jockeying for space on a dance floor or two armies jockeying for a truce line may jointly suffer from decision processes that are limited to the abstract properties of the situation.

A particular implication of this general point is that the game in "normal" (mathematically abstract) form is not logically equivalent to the game in "extensive" (particular) form, once we admit the logic by which rational players concert their expectations of each other. It will be argued later, and is also argued in the author's earlier paper in this *Journal* (31), that these same considerations are powerfully present in explicit bargaining as well. A terminological implication of these considerations is that "non-co-operative" is a poor name for the game of tacit co-ordination; it is desperately co-operative in its own peculiar way and is still so when we add conflict and form the tacit mixed-motive game. (In separate papers the writer has attempted to demonstrate that certain solution concepts familiar in game theory can be given an interpretation in terms of the co-ordination concept [32, 33].)

SUGGESTION AND MUTUAL PERCEPTION IN THE MIXED-MOTIVE GAME

Co-ordination-game theory, while interesting in its own right, is interesting mainly for the light that it sheds on the nature of the mixed-motive game. The co-ordination element shows up most strikingly in a purely tacit game, in which there is neither communication nor any sequence of moves by which the two players accommodate themselves to each other. An example, similar to one of those reported in the earlier article, would be the following.

One player is "located" in Cincinnati, the other in San Francisco; they have identical maps of the United States and are to divide the country between them. Each is to draw a line dividing the United States into two parts; the line may be straight or curved, related or unrelated to physical or political landmarks. If the two of them divide the map differently, neither gets anything; but if they draw identical division lines on their maps, they are both rewarded. The reward for each player depends on what is contained in his piece after the division, i.e., the piece that contains the city in which he is located. Let us leave these rewards vague; they may depend partly on area, partly on population, partly on industrial wealth and agricultural resources, etc., and may differ somewhat for the two players. In other words, while all terrain is valuable, not all parts of the country are equally valuable, and there is no clear specification of the valuation formula. (There is consequently no means of selecting a perfectly symmetrical division of values between the two players.)

In this game there is a compelling problem of co-ordination; each player can win only if he does exactly what the other expects him to, knowing that the other is similarly trying to do exactly what is expected of him. They must jointly find a line that in some fashion suggests itself to both of them or appeals to both of them. Neither can "outsmart" the other without outsmarting himself.

The experiments reported earlier suggest that players are by no means helpless when faced with this kind of game. The game is nowhere near so "infinitely" difficult as the infinity of possible division lines might suggest; some variants of the game are not difficult at all. But a successful outcome does depend on the kinds of factors that are controlling in the pure-co-ordination game; in fact, some games of this sort are "won" by the two players' choosing exactly the same outcome as they would have chosen if the reward system gave them

identical, instead of conflicting, interests. The problem is to find some signal or clue or rationalization that both can perceive as the "right" one, with each party prepared to be disciplined by that signal or clue in the event that it appears to discriminate against him. They must find their clues where they can. (If the map they are using happens, for example, to contain an embarrassing richness of clues, making it difficult to single out any particular one, a fairly arbitrary line drawn as a suggestion by the referee, identical on both maps, might have to be accepted as a "mediator," even if it is substantially biased toward one of the players.)

But this co-ordination element, especially in the case without conflict, appears to be essentially related to a *communication problem.* The pure-co-ordination game not only ceases to be interesting but virtually ceases to be a "game" if the players can concert with certainty, without difficulty, and without cost. The question arises, then, how important the co-ordination element can be in mixed-motive games generally, since many of these take the form of overt bargaining with uninhibited speech.

The pervasiveness of the co-ordination principle arises from two separate considerations. One, which will be elaborated in Part III of this paper, is that tacit bargaining provides an analytical model—perhaps only an analogy but perhaps an identification of the actual psychic and intellectual phenomenon—of the "rational" process of finding agreement in pure bargaining situations, those in which both parties recognize that there is a wide range of outcomes preferable to both of them over no agreement at all. The psychic phenomenon of "mutual perception" that can be verified as real and important in the tacit case has a role to play in the analysis of explicit bargaining. (*Co-ordination of expectations* will be the related concept.)

Second, many of the bargaining processes or game situations that we want to analyze are at least partly tacit. In some cases, like maneuvering a car in a traffic jam, speech is physically precluded; in others, like developing a modus vivendi with a neighbor, speech is inhibited in the interests of privacy. Illicit bargaining or diplomatic bargaining that would be embarrassing to both sides if overheard by other countries may be less than fully articulate. If the number of players in a game is large, as it is in the bargaining process that determines the racial border lines between residential areas and professions, there may be no institutional provision for explicit negotiation. In these cases, while speech may be part of the bargaining process, actions are also part of it, and the game is one of "maneuver" rather than just talk.

Furthermore, if there are *moves* available to the players, so that it is an advantage to get on with the maneuver even while negotiating, and particularly if some maneuvers become visible to the other player only after a time lag, there is no reason to suppose that an instantaneous moritorium on maneuver will reign from the outset; in that case, the game progresses while the talk is going on. If the moves had only symbolic significance, we could include them in the communication process along with speech; but, typically, moves have a tactical significance, leaving the game irreversibly different from what it was before, and typically also their tactical significance raises them above the level of pure speech even in their communication content. One may say and say that a gun is loaded without being able to prove it until he actually shoots; one may say and say that he considers an area strategically important

and not be believed until he incurs expense or risk in its protection. Thus moves can reveal information about a player's value system or about the choices of action available to him; moves can commit him to certain actions when speech often cannot; and moves can often progress at a speed that is determined unilaterally, not dependent on formalities of agreement at a conference.

In other words, bargaining games quite typically involve a dynamic process of mutual accommodation rather than pure communication culminating in a crystallized agreement. The jockeying for limits in limited war is a perfect example, and we might illustrate it by modifying the parlor game described above.

An illustrative tacit game. Suppose our two players with their maps of the United States before them are each given 100 chips and told to play a game as follows.[12] At each "move," each player will distribute five chips among states on his map. The moves are compared, and if the two players have put a chip apiece in the same state, those two chips are removed; if one player has put a chip and the other player three chips in the same state, a chip apiece is removed leaving only two chips representing the one player; etc. They do the same at the next move, again with five chips; this time they have the option of placing their chips on states that are yet uncovered or of placing them on states where there are already chips. If A puts two chips on a state in which B previously put a chip, B's is removed along with one of A's, leaving one of A's chips present to "claim" the state. And so the game goes until the players have used up all their chips; it then continues, and at each move a player may transfer up to five chips from the states in which they are to other states, again with equal numbers of chips being removed from a state in which both players have placed chips. This process goes on until both players have notified the referee that they are willing to terminate the game.

Prizes are now distributed. Each player receives a dollar for every one of his chips still on the board, i.e., for those that were not removed when he "took" a state or "lost" it to the other player. He also gets money for the states that he "possesses," these being the states that he has chips on plus those without chips that are in the area containing his home base that is completely inclosed by states that he does have chips on.

These "rewards" for states possessed are specific dollar values attached to each of the 48 states; they vaguely follow a pattern suggestive of, say, "economic worth" or something of the sort. There is no presumption that the values are the same, or even very closely correlated, for the two players; population may be an important

[12] Since it will be proposed later in this paper that such games have, in fact, a research value, as well as an illustrative value, it should be observed at the outset that there is a special problem of motivating the players in an experimental non-zero-sum game. In a zero-sum game, winning is measured relative to one's immediate adversary, and the intellectual challenge and bilateral competition motivate the player toward the correct (and only) type of winning. But for a mixed-motive game, "winning" must be made to involve one's absolute score, not his score relative to that of the person he plays with; the incentives are distorted if the play is dominated by strictly bilateral competition. So, unless real rewards are given, the game has to be organized as a round robin or some such schedule that involves more than two players in a series of two-person plays, with the final outcome decided by the relative position of one's absolute score. (This is why there are no two-person non-zero-sum parlor games.)

element in the "values" of the states for one of the players and a comparatively unimportant element in the "values" for the other player. Neither player knows the other player's value system—or perhaps knows just a little about it, such as what elements matter but not how much they matter. Each must learn what he can about the other's value system by observing the other player's moves.

Here we have a mixed-motive game, which progresses by a process of mutual accommodation—a series of moves in the course of which the players suffer damage jointly if their accommodation is poor. They may lose dollars by failing to predict where each other will place his chips during the current move, in those cases where they prefer not to lose dollars fighting over a state. Each loses at least a dollar when one takes a state from the other; and they may lose more than a dollar apiece if the one who loses a state attempts to recapture it by putting more chips on it. And not only do they lose a dollar with each dollar forfeited, but each player has fewer "chips" left from the point of view of claiming states; and they may have to leave some states completely unclaimed between them if they have not enough chips left on the board when the game ends.

Now how do the players "bargain" in this game? One way or another, they do in fact make proposals and counterproposals; they accept, reject, retaliate, and even discover ways of conveying threats and promises. But if we deny them any form of speech, they must convey their intentions and their proposals by their patterns of behavior. Each must be alert to what the other is expressing in his maneuvers, and each must be inventive enough to convey his intentions when he wants them conveyed. If one player badly wants a particular state, because it has especially high value for him, so that he is willing to stick around and fight it out a long time, losing several dollars to the kitty before the other player gives up, it is better for both players that they realize ahead of time which one wants it most badly. And if a player is really prepared to concede a large portion of the country as a "trade" for some other portion that he badly wants, he must not only make it conspicuously available to the other side but must somehow demarcate its limits by his own pattern of play.

But where do the patterns come from? They are not very richly provided by the mathematical structure of the game, particularly since we have purposely made each player's value system too uncertain to the other to make considerations of symmetry, equality, etc., of any great help. Presumably, they find their patterns in such things as natural boundaries, familiar political groupings, the economic characteristics of states that might enter their value systems, Gestalt psychology, and any clichés or traditions that they can work out for themselves in the process of play.[13]

Explicit communication. Now let us change the rules so that the players may talk as much as they please. How different would this make the game? In some respects, it should increase the efficiency of the players; particular trades can be identi-

[13] If my neighbor's fruit tree overhangs my yard and I pick exactly all the fruit on my side of the line, my neighbor can probably discern what my "proposal" is, and has a good idea of what he has acquiesced in for the future if he does not retaliate. But if, instead, I pick that same amount of fruit from both sides of the line haphazardly or pick some amount that is related, say, to the size of my family, he is less likely to perceive just what I have in mind. (He may also be more obliged to resist or retaliate if I pick only *part* of the fruit on my side of the line than if I pick it all, since I have failed to demarcate the limit of my intentions.)

fied now that were too complex to make proposals about under the more clumsy system. Perhaps, too, the players can avoid some of the inadvertent clashes of chips on the same state, which cost them dollars. We cannot be sure that they will avoid mutually costly competitive bidding for states, since the advantage of being first on a state is great enough to motivate players to keep playing even while they talk. And they have no way to persuade each other that they mean what they say except by showing it in the way they play. (We let them tell each other how they value the states; but we explicitly make fibs unpunishable, and we provide the players no written evidence of their value systems that they could show each other.)

So the introduction of uninhibited speech may not greatly alter the character of the game, even though the particular outcome is different. The dependence of the two players on conveying their intentions to each other and perceiving the intentions of each other, of behaving in predictable patterns and acquiescing in rules or limits, is much the same as before.

The contrast with a zero-sum game and the peculiarly self-effacing quality of a minimax solution is striking here. With a minimax solution, a zero-sum game is reduced to a completely unilateral affair. One not only does not need to communicate with his opponent, he does not even need to know who the opponent is or whether there is one. A randomized strategy is dramatically anticommunicative; it is a deliberate means of destroying any possibility of communication, especially communication of intentions, inadvertent or otherwise. It is a means of expunging from the game all details except the mathematical structure of the payoff, and from the players all communicative relationships.

In chess it does not matter whether the pieces look like horses, ecclesiastics, elephants, castles, or hamburger buns; whether the game is called "chess," "civil war," or "real estate"; or whether the squares are distorted to look like political or geographical subdivisions. It does not matter what the players know about each other or whether they speak the same language and have a common culture; nor does it matter who played the game previously and how it came out. (If it did matter, one of the players would be motivated to destroy the influence of these details; and a minimax strategy, randomized if necessary, would destroy it.)

But change the payoff matrix in a chess game, making it a non-zero-sum game that rewards the players not only for the pieces they capture but for the pieces they have left over at the end, as well as the squares they occupy, in such fashion that both players have some interest in minimizing the "gross" capture of pieces with its mutual destruction of value. Make each player uncertain about just what squares and what particular pieces the other player values most. And have moves by the clock, so that neither player can hold up the other player's moves for the sake of talking to him.

Now it may make a difference to the players whether we call the game "war" or "gold rush"; whether the pieces look like horses, soldiers, explorers, or children on an Easter egg hunt; what map or picture is superimposed on the playing board and how the squares are distorted into different shapes; or what background story the players are told before they begin.

We have now rigged the game so that the players must *bargain* their way to an outcome, either vocally or by the successive moves that they make, or both. They must find ways of regulating their behavior, communicating their intentions, letting themselves be led to some meeting of minds,

tacit or explicit, to avoid mutual destruction of potential gains. The "incidental" details may facilitate the players' discovery of expressive behavior patterns; and the extent to which the *symbolic* contents of the game—the suggestions and connotations —suggest compromises, limits, and regulations should be expected to make a difference. It should, because it can be a help to both players not to limit themselves to the abstract structure of the game in their search for stable, mutually non-destructive, recognizable patterns of movement. The fundamental psychic and intellectual process is that of participating in the creation of *traditions;* and the ingredients out of which traditions can be created, or the materials in which potential traditions can be perceived and jointly recognized, are not at all coincident with the mathematical contents of the game.[14]

The outcome is determined by the expectations that each player forms of how the other will play, where each of them knows that their expectations are substantially reciprocal. The players must jointly discover and mutually acquiesce in an outcome or in a mode of play that makes the outcome determinate. They must together find "rules of the game" or together suffer the consequences.

A good example of this problem of communicating intentions is that of getting across, persuasively, an intended pattern of retaliation for particular acts that one proposes to consider "out of bounds." Without full communication, one's ability to convey such a pattern of intentions is dependent not only on the contextual materials available for the formation of bounds and limits but on the capacity of the other player to recognize the formula (Gestalt) of retaliation when he sees a sample of it. Historical and literary precedent, legal and moral casuistry, mathematics and aesthetics, as well as familiar analogues from other walks of life, may constitute the menu from which one has to choose his recognizable pattern of retaliation as well as his interpretation of the other's intended pattern. Even with full verbal communication, the situation may not be greatly different; patterns of action may speak louder than words.

Thus the influence that the suggestive details of a game may have on its outcome and the dependence of the players on what clues and signals the game provides are relevant not merely to the study of how players actually do behave in a non-zero-sum game. It is not being argued that players just *do* respond to the non-mathematical properties of the game but that they *ought* to take them into account, hence that even a normative theory—a theory of the *strategy* of games—must recognize that rational players may jointly take advantage of them. And even when one rational player

[14] A good example is the question whether a clear line can be drawn between atomic and other weapons, the answer to which is reported now to be negative if explosive power is the criterion, the explosive ranges having overlapped. But there is nevertheless a difference if enough people think so, and they undoubtedly do. It is a difference constructed of the pure fabric of expectations: it is a ten years' *tradition* that atomic weapons *are* different; people believe so and believe others to believe so, and even those who deny the difference will undoubtedly catch their breath, whenever the next one goes off in a war, in a manner they cannot explain by reference to the force of the explosion. It is a purely conventional difference, like the one that makes imprisonment not a "cruel and unusual" punishment or that makes, say, university representation in Parliament perfectly compatible with English democracy if it has always existed but not if it has to be reinstated after a ten years' lapse. The atomic-weapons difference is also one that, probably, can be deliberately reinforced or deliberately blurred over time, as most traditions can.

realizes that the configuration of these details discriminates against him, he may also rationally recognize that he has no recourse —that the other player will rationally expect him to submit to the discipline of the suggestions that emanate from the game's concrete details and will take actions that—on pain of mutual damage—assume he will co-operate.

Communicating subjective information. The role of "expressive moves" in a mutual-accommodation game of this sort is enhanced by the consideration that in mixed-motive games, in contrast to zero-sum-games that are known to the players to be zero-sum, there is likely to be uncertainty about each other's value system. Moves have an *information* content in the mixed-motive game.

Nor can we set up as a general case the bargaining game in which each side has foreknowledge of the other's preferences. To assume that either knows the "true" payoff matrix of the other is often to make an extraordinary assumption about the institutional arrangements of the game. The reason is that certain elements in a bargaining game are *inherently unknowable* for some of the participants, except when there are special conditions. How can we know how badly the Russians would dislike an all-out war in which both sides were annihilated? We cannot; and the reason we cannot is *not* solely that the Russians are necessarily unwilling that we should know. On the contrary, circumstances may arise in which they are desperate that we should know the truth. But how can they make us know it? How can they make us believe that what they tell us is true? How can the prisoner being tortured for secrets that he really does not know persuade his captors that he does not know them? How could the Chinese, if they were really determined to take Formosa at the cost of an all-out war, persuade us that they could not be deterred in any fashion and that any threat on our part would only commit us both to all-out war?[15]

In special cases the information can be conveyed. In an artificial game, in which each player's "value system" is contained on cards or chips, he may simply turn them face up (if the rules permit or if he and his adversary can jointly cheat against the referee). In a society that believes absolutely in a superior power that will punish falsehood when asked to do so and that everybody knows everybody else believes in, "cross my heart and hope to die" is a sufficient formula for conveying truth voluntarily. But these are special cases. If we are to have a "general case" it must be one in which there is at least some ignorance of each other's value system, or each other's strategy choices, if only because such facts are inherently unknowable or incommunicable.

Von Neumann and Morgenstern (38) illustrated their *solution* concept for the non-zero-sum game with the example of a seller,

[15] The lack of any means of testing the truth is the very basis of that tantalizing game in which each participant attaches positive value to the other's welfare, as when husband and wife discuss whether or not to go to a movie, each wanting to do whatever the other wants to do and wanting to seem to want it himself, knowing that the other is similarly expressing a preference that represents a guess at what one wants to do, etc. There is also an entire domain of game theory involving interpersonal relations in which the overt revelation or recognition of one's value system itself affects values; my awareness that my neighbor does not like me may cause me small discomfort, as does his awareness of my awareness, but if we are forced to accredit the fact overtly, the pain may be acute. "Social etiquette," remarks Erving Goffman (11, p. 224), "warns men against asking for New Year's Eve dates too early in the season, lest the girl find it difficult to provide a gentle excuse for refusing."

A, prepared to sell his house for any price above 10, and two buyers prepared to pay up to 15 and 25, respectively. (My numbers.) The novel part of the solution was that C might pay B a share of his saving if, through B's staying out of the market, C got the house for less than 15. They proposed—and this limitation was inherent in their concept of *solution*—that the most B might receive from C was $15 - 10 = 5$. What is interesting about the information requirement of this solution is not that B's reservation price of 15 is something that he might try to misrepresent but that in the ordinary world he could not convincingly communicate the truth if he wanted to. Not only does the "solution" concept—by its assumption of full information—rule out the intrusion of speculators (unless they genuinely want the house enough to give them a basis for sharing in the solution), but it assumes that C can discern, or B can reveal, a subjective truth, one that D and E (speculators who are attracted by the observation that B makes a pure bargaining profit in connection with an object that he never owns before or after) cannot counterfeit.

There are undoubtedly special cases in which one can suppose that the other player is like one's self in basic values and can consequently estimate the other's values by the simple application of symmetry. But in too many exciting cases one plays an opponent who is a wholly different kind of person. The father of a kidnapped boy will not be very successful in guessing what his own bottom price would be if he had been the kidnapper instead; it may not be easy for a British or French officer introspectively to guess how terrible a penalty would have to be to deter him if he were a Mau Mau or an Algerian terrorist. It is hard for a boy to guess how much he would like himself if he were the girl that he wants to date, or for the customer in the restaurant to know how much he would dislike a scene if he were the waiter instead.

This is one of the reasons why talk is not a substitute for moves. Moves can in some way alter the game, by incurring manifest costs, risks, or a reduced range of subsequent choice; they have an information content, or *evidence* content, of a different character from that of speech. Talk can be cheap when moves are not (except for the "talk" that takes the form of *enforcible* threats, promises, commitments, and so forth, and that is to be analyzed under the heading of *moves* rather than communication anyway). Mutual accommodation ultimately requires, if the outcome is to be efficient, that the division of gains be in accordance with "comparative advantage"; that is, the things a player concedes should be those that he wants less than the other player, relative to the things he trades for. Each needs, therefore, to communicate his value system with some truth, although each can also gain by deceiving. While one's maneuvers are not unambiguous in their revelation of one's value systems and may even be deliberately deceptive, they nevertheless have an evidential quality that mere speech has not.

The uncertainty that can usually be presumed to exist about each other's value systems also reduces the usefulness of the concept of mathematical *symmetry* as a normative or predictive principle. Mathematical symmetry cannot be perceived if one has access to only half the relevant magnitudes. To the extent that symmetry is helpful to the players in accommodating their movements to each other's, it would tend to be symmetry of a more qualitative sort, of the kind that depends on visible context rather than underlying values.

II. *Enforcement, Communication, and Strategic Moves*

Whenever we speak of deterrence, atomic blackmail, the balance of terror, or an open-skies arrangement to reduce the fear of surprise attack; when we characterize American troops in Europe as a trip wire or plate-glass window or propose that a threatened enemy be provided a face-saving exit; when we advert to the impotence of a threat that is so enormous that the threatener would obviously shrink from carrying it out or observe that taxi drivers are given a wide berth because they are known to be indifferent to dents and scratches, we are evidently deep in game theory. Yet formal game theory has contributed little to the clarification of these ideas. The author suggests that non-zero-sum game theory may have missed its most promising field by being pitched at too abstract a level of analysis. By abstracting from communication and enforcement systems and by treating perfect symmetry between players as the general case rather than a special one, game theory may have overshot the level at which the most fruitful work could be done and have defined away some of the essential ingredients of typical non-zero-sum games. Preoccupied with the solution to *the* non-zero-sum game, game theory has not done justice to some typical game situations or game models and to the "moves" that are peculiar to non-zero-sum games of strategy.

What "model," for example, epitomizes the controversy over massive retaliation? What conditions are necessary for an efficacious threat? What in game theory corresponds to the proverbial situation "to have a bear by the tail"; how do we identify the payoff matrix, the communication system, and the enforcement system that it embodies? What are the tactics by which pedestrians intimidate automobile drivers, or small countries large ones; and how do we formulate them in game-theoretical terms? What is the information or communication structure, or the complex of incentives, that makes dogs, idiots, small children, fanatics, and martyrs immune to threats?

The precarious strategy of cold war and nuclear stalemate has often been expressed in game-type analogies: two enemies within reach of each other's poison arrows on opposite sides of a canyon, the poison so slow that either could shoot the other before he died; a shepherd who has chased a wolf into a corner where it has no choice but to fight, the shepherd unwilling to turn his back on the beast; a pursuer armed only with a hand grenade who inadvertently gets too close to his victim and dares not use his weapon; two neighbors, each controlling dynamite in the other's basement, trying to find mutual security through some arrangement of electric switches and detonators. If we can analyze the structures of these games and develop a working acquaintance with standard models, we may provide insight into real problems by the use of our theory.

To illustrate, an instructive model is that of 20 men held up for robbery or ransom by a single man who has a gun and six bullets. They can overwhelm him if they are willing to lose six of themselves, if they have a means of deciding which six to lose. They can defeat him without loss if they can visibly commit themselves to a threat to do so, if they can simultaneously commit themselves to a *promise* to abstain from capital punishment, once they have caught him. He can deter their threat if he can visibly commit himself to shoot in disregard of any subsequent threat they might make, or if he can show that he could not believe

their promise. If they cannot deliver their threat—if, say, he understands only a foreign language—they cannot disarm him verbally. Nor can they make a threat unless they agree on it themselves; so, if he can threaten to shoot any two who talk together, he can deter agreement. If the 20 cannot find a way to divide the risk, there may be no one to go first to carry out the threat, hence no way to make the threat persuasive; and if he can announce a formula for shooting, such as that those who move first get shot first, he can deter them unless they find a way to move together without a "first." If 14 of the 20 can overpower the remaining 6 and force them to advance, they can demonstrate that they could overwhelm the man; if so, the threat succeeds, and even the 6 "expendables" gain through their own inability to avoid jeopardy. If the 20 could overwhelm the man but have no way of letting him escape, a promise of immunity may be necessary; but if they cannot deny their capacity to identify him and testify against him later, it may be necessary to let him take a hostage. This, in turn, depends on the ability of 19 to enforce their own agreement to protect, by silence, whoever is currently the hostage. . . . And so on. When we have identified the critical ingredients in several games of this sort, we may be in a better position to understand the basis of power of an unpopular despot or of a well-organized dominant minority, or the conditions for successful mutiny.

This section (Part II) of the paper is an attempt to suggest the kinds of typical moves and structural elements that deserve to be explored within the framework of game theory. They include such moves as "threat," "promise," "destruction of communication," "delegation of decision," and so forth, and such structural elements as the communication and enforcement provisions.

AN ILLUSTRATIVE MOVE

As an example of a standard "move," consider the bargaining between a person who has a particular object of art and someone else who is the only person interested in buying it, and suppose that each has a pretty good idea how badly the other wants it. The two are haggling over a price. We have not yet defined a game: "haggling" is not a well-defined concept. But suppose the institutional environment makes it possible for the potential buyer, who knows approximately the minimum price at which the potential seller would part with it, to make a single "final" offer subject to extreme penalty in the event that he should amend the offer. This possibility of invoking a penalty permits the potential buyer to commit himself, in a way that irreversibly changes the character of the game. There remains but a single decision for the potential seller: to sell at the price proposed or to keep the object indefinitely. The possibility of commitment converts an indeterminate bargaining process into a two-move game; one player assumes a commitment, and the other makes a final decision. The game has become determinate.[16]

[16] In the real estate example of Von Neumann and Morgenstern referred to earlier (p. 219) buyer B (whose top price is 15) might raise the limit on what he can extract from buyer C (whose top price is 25) if he can find some means to bind himself to buy the house for 20 and keep or destroy it (i.e., not be free to resell it to C for a loss) unless he gets a specified large fraction of, say, $20 - P$, where P is the ultimate price paid by C. In effect, B changes his own "true" top price, thus raising the limit on what he may extract from C. Of course, D and E may try to do the same; and the first to get properly committed, or the one who can find a means if only one of them can, is the winner. If D, who attaches no personal value to the house, is committed to pay up to 22 for it, he is a bona fide member of the game with a true reservation price of 22; his *bona fides* is

This particular move was analyzed at some length in an earlier article (30) and is mentioned here only as a particularly simple illustration of a typical move. Certain points should be noted about it. First, the availability and the efficacy of this move depend on the communication structure of the game and the ability of the player to find a way to commit himself, to "enforce" the commitment against himself. Second, we have allowed the move structure of the game to be asymmetrical; the "winner" is the one who can assume the commitment or, if both can, the one who can do it first. (We can consider the special case of a tie, but we have not, by an assumption of symmetry, made ties a foregone conclusion.)

Third, although we have made the game "determinate" in the sense that we have no difficulty in identifying the "solution," once we have identified which of the two players can first commit himself, it remains a game of *strategy*. Though the winner is the one who achieves his commitment first, the game is not like a foot race that goes to the fastest. The difference is that the commitment does not automatically win under the rules of the game, either physically or legally. The outcome still depends on the second player, over whom the first player has no direct control. The commitment is a *strategic* move, a move that induces the other player to choose in one's favor. It constrains the other player's choice by affecting his expectations.

The power to commit one's self in this kind of game is equivalent to "first move." And if the institutional arrangements provide no means for incurring an irrevocable commitment in a legal or contractual sense, one may accomplish the same thing by an irreversible maneuver that reduces his own freedom of choice. One escapes an undesired invitation by commitment when he arranges a "prior" engagement; failing that, he can deliberately catch cold. Luce and Raiffa (21, p. 75) have pointed out that the same tactic can be used by a person against himself when he wants, for example, to go on a diet but does not trust himself. "He announces his intention, or accepts a wager that he will not break his diet, so that later he will *not* be free to change his mind and to optimize his actions according to his tastes at *that* time." The same thing is accomplished by maneuver rather than by commitment when one deliberately embarks on a vacation deep in the wilds without cigarettes.

THREATS

The distinctive character of a threat is that one asserts that he will do, in a contingency, what he would manifestly prefer not to do if the contingency occurred. The motive behind the threat is to coerce or to deter, to constrain the other player's choice of action. It works by altering the other player's expectation of how the threatening player would react; it is an attempt to alter the threatener's own *incentives* as they are seen by the party threatened.

One threatens retaliation not because, if the provocative act occurs, one has anything to gain by retaliating; instead one risks having to retaliate in the hope that, by the sheer act of creating the risk, he will deter the act that retaliation is made contingent on. There is consequently a motive to make the threat but not to carry it out. More correctly, there is a motive to bind one's self so that fulfilling the threat is obligatory; but if the threat fails, so that it has to be carried out, the only motive for carrying it out is the obligation that was deliberately incurred earlier (plus any motive arising from the

---

even greater than was B's originally, if the commitment is demonstrable while subjective valuations are not.

likelihood that fulfilment in this case increases the potency of some future threat or other simultaneous threat—i.e., from the evidence-value of the fulfilling action).

The threat is related to the commitment in two ways. First, like the commitment, it is a surrender of choice, a renunciation of alternatives, that makes one worse off than he need be in the event that the tactic fails; the threat and the commitment are both motivated by the possibility that a rational second player can be constrained by his knowledge that the first player has altered his own incentive structure. Both tactics are intended to rig one's own incentive structure so that the other player is left the initiative and will be induced to choose in the first player's favor.

Second, the threat is related to the commitment in that it depends on it; the threat can constrain the other player only insofar as it carries to the other player at least some appearance of obligation. If one is not committed to the threat in any way and cannot even seem to be committed, it is ineffectual. If I threaten to blow us both to bits unless you close the window, you know that I won't unless I have somehow managed to leave myself no choice in the matter.

The threat differs from the commitment, however, in that it makes one's courses of action *conditional* on what the other player does at his next turn; while the commitment fixes one's courses of action, the threat fixes a course of reaction, of response to the other player. The commitment is a means of gaining *first move* in a game in which *first move* carries an advantage; the threat is a commitment to a strategy for *second move.*

A threat can be effective only if the game is one in which the first move is up to the other player or one can force the other to move first. But if one must, in a mechanical sense, move first or simultaneously, he can still force the legal equivalent of "first move" on the other by attaching his threat to a demand that the other promise in advance how he will behave—if the game has communication and enforcement structures that make promises feasible and that the party to be threatened cannot destroy in advance. The holdup man whose rich victim happens to have no money on him at the time can make nothing of his opportunity unless he can extract a hostage while he awaits payment; and even that will not work unless he can himself find a way to assume a convincing commitment to return the hostage in a manner that does not subject himself to identification or capture.

The fact that *some* kind of commitment, or at least appearance of commitment, must lie behind the threat and be successfully communicated to the threatened party is in contradiction to another notion that often appears in game theory. This is the notion that a threat is desirable, or admissible, or plausible, only if the reaction threatened would cause worse damage to the threatened party than to the party making the threat. This is the view of Luce and Raiffa (21), who characterize threats by the phrase, "This will hurt you more than it hurts me," explicitly making threats depend on interpersonal utility comparisons. In the event that both players attempt to make plausible threats, they say, the result becomes indeterminate, depending on the "bargaining personalities" of the players; "and to predict what will in fact happen without first having a complete psychological and economic analysis of the players seems foolish indeed."[17]

[17] Morton A. Kaplan (15, p. 198), in applying game theory to international relations, also takes the position that "any criterion giving weight to the threat positions of the players involves an interpersonal comparison of utilities." Luce and Raiffa (21, pp. 110–11, 119–20, 143–44) may be led partly to their view that only one of the

But the issue is both simpler and more precise than that. Consider the left-hand matrix in Figure 2, where Column is assumed to have "first move." Without threats, Column has an easy "win." He chooses strategy I, forcing Row to choose between payoffs of 1 and 0; Row chooses strategy i, providing Column a payoff of 2. But if we allow Row to make a threat, he declares that he will choose strategy ii unless Column chooses II; that is, he gives Column a choice of ii,I or i,II by committing himself to that conditional choice. *If* Column went ahead and chose I, of course, Row would prefer to choose i; and they both know it. The tactic succeeds only if Column believes that Row *must* choose ii in the event of I.

Either he does believe this, or he does not. If he does not, the "threat" is nothing at all to him; he goes ahead and makes his "best" first move, choosing I. If he does believe that Row must follow a strategy of i,II

	I	II
i	2 1	1 2
ii	0 0	0 0

	I	II
i	10 9	9 10
ii	8 0	0 0

FIG. 2

players has a "plausible" threat to make, by confining their brief discussion to 2×2 matrices. It is impossible to show, with a 2×2 matrix, a game in which both players could be interested in making threats. A threat is essentially a credible declaration of a *conditional* choice for second move. It is profitable only if it yields a better payoff than either first move or second move alone and when one can make the other player move first either actually or by promise. (If second move alone is as good, the threat is unnecessary; and if first move were as good, one needs only an unconditional commitment to his strategy choice, not a commitment to a conditional choice.) But if this preference order holds for one player in a 2×2 matrix, it cannot hold for the other player. The actual matrices used by Luce and Raiffa in discussing the point show no "plausible" threat strategy for player No. 2, not because the absolute size of his gains or losses is greater than player 1's but for the much simpler reason that player 2 has no use for a threat. He wins if he moves first; he wins if he moves second; and he wins with simultaneous moves, in the games shown. His only interest in a threatlike declaration would be to forestall his partner's threat; and for that purpose he needs only an *unconditional* commitment to his preferred strategy—i.e., the legal equivalent of "first move" in advance of his partner's threat. The "threat tactic" of J. F. Nash (27), which applies to bargaining games that have a continuous range of efficient outcomes—or that can be made to, by agreement on the odds in a drawing of lots —differs from the threat discussed here, in that the threatener does not demand, on pain of mutual damage, a *particular* outcome but only *some* outcome in the efficient range; that is, he shifts the zero point corresponding to "no agreement." The motive for that threat is the expectation of a particular mathematically determinate outcome whose locus is shifted by the shift in the payoffs corresponding to non-agreement. (For further discussion of Nash's theory see Harsanyi [13].) This is the kind of threat assumed by Luce and Raiffa (21, p. 139) in the "asymmetrical" game. The implicit legal structure of the game apparently honors no irrevocable commitments (otherwise, first commitment would easily win the game for either player). Each player is subject to the legal "disability" that he can always, by the overt act of explicit agreement with his partner on any outcome, evade his own commitment. This being so, the revocable commitments can only shift the zero point—the "status quo" that will rule unless explicit agreement on some outcome is reached. The "asymmetry" that is present in the particular game shown by Luce and Raiffa is thus a feature of the particular legal system that implicitly prevails. In practice it might correspond, say, to the deliberate incurring of social disapproval on failure to reach agreement, with such disapproval constituting cost or punishment (perhaps asymmetrical between participants) in addition to the cost of non-agreement but with the public not concerned with what the agreement provides as long as some agreement is reached.

or ii,I, Column prefers 1 to 0 and chooses II. But this is true of any numbers that we might put in the matrix that reflect the same order of preferences. It is true of the right-hand matrix as well. That one dramatizes the essential character of the threat more than the first one, since the penalty on Row of an irrational choice by Column is greater in this case; but for rational play and full information, Row need not worry. Column's preference is clear; and, once Row has given him the pair to choose from—ii,I versus i,II —there is no doubt what Column will do. If I threaten to blow my own brains all over your new suit unless you give me that last slice of toast, you'll give me the toast or not depending on whether you know that I've arranged to have to do so, exactly as if I'd only threatened to throw my scrambled eggs at you.

The issue here is in whether or not we admit that the game has "moves," i.e., that it is possible for one player or both players to take actions in the course of the game that irreversibly change the game itself—that in some fashion alter the payoff matrix, the order of choices, or the information structure of the game. If the game by its definition admits no moves of any sort, except mutual agreement and refusal to agree, then it may be true that the "personalities" of the players determine the outcome, in the sense that their expectations in a "moveless" game converge by a process that is wholly psychic. But if a threat is anything more than an assertion that is intended to appeal to the other player by power of suggestion, we must ask what more it can be. And it must involve some notion of commitment—real or fake—if it is to be anything.

"Commitment" is to be interpreted broadly here. It includes maneuvers that leave one in such a position that the option of non-fulfilment no longer exists (as when one intimidates the other car by driving too fast to stop in time), maneuvers that shift the final decision beyond recall to another party whose incentive structure would provide an ex post motive for fulfilment (as when the authority to punish is deliberately given to sadists, or when one shifts his claims and liabilities to an insurance company), and maneuvers that simply "worsen" one's own payoff in the contingency of non-fulfilment so that even the horror of a mutually damaging fulfilment becomes more attractive (as when one arranges for himself to appear a public coward if he fails to fulfil, or when he puts a plate-glass window in front of his wares or stations women and children on the particular bit of territory that he has threatened somewhat implausibly to defend at great cost). A nice everyday example is given by Erving Goffman (11, p. 215), who reminds us that "Salesmen, especially street 'stemmers,' know that if they take a line that will be discredited unless the reluctant customer buys, the customer may be trapped by considerateness and buy in order to save the face of the salesman and prevent what would ordinarily result in a scene."[18]

There are, however, some ways in which this notion of commitment to a threat can be usefully loosened. One is to recognize that "firm" commitment amounts to the invocation of some wholly potent penalty, such that one would in all circumstances prefer to carry out what he was committed to. It is a penalty of infinite (or at least of superfluous) size that one voluntarily, irreversibly, and visibly attaches to all patterns of action but the one that he is committed to. This concept can be loosened by supposing that the penalty is of finite size and not nec-

[18] Goffman's paper is a brilliant study in the relation of game theory to gamesmanship and a pioneer illustration of the rich game-theoretic content of formalized behavior structures like etiquette, chivalry, diplomatic practice, and—by implication—public law.

essarily so large as to be controlling in all cases. In Figure 3, Column will win if he has first move, unless Row can commit himself to i. (Commitment obtains "first move" for Row.) But if commitment means the attachment of a finite penalty to the choice of row ii and we show this in the matrix by subtracting from each of Row's payoffs in ii some finite amount representing the penalty, then the commitment will be effective only if the penalty is greater than 2. Otherwise it is clear to Column that Row's response to II will be ii, in spite of the commitment. In this case the commitment is simply a loss that Row would impose on himself, so he avoids it.

Similarly with a threat. In Figure 4, without threats, the solution is at iii,II whether the rules call for Row to choose first, Column first, or both to choose simultaneously. Either player can win if he can move second and confront the other with a threat.[19] Column would threaten I against iii, Row would threaten i against II. But if the threat is secured by a penalty, the lower limit to any persuasive penalty that Column could invoke would be 4; any smaller penalty leaves him preferring II to I when Row chooses iii. The lower limit to a persuasive penalty on Row's non-compliance would be 3. If, then, the situation is one in which penalties come in a single "size," a size less than 3 goes unused and the outcome is at iii,II; a size greater than 4 is adequate for either player, and the "winner" is the one who can avail himself of the threat first; a size between 3 and 4 is of use only to Row,

	I	II
i	2 4	1 1
ii	2 4	3 3

FIG. 3

	I	II	III
i	−5 −5	−1 −2	−1 −2
ii	−3 −4	3 0	2 2
iii	−3 −4	1 1	0 3

FIG. 4

who wins. In this latter case the player who would be hurt the more by his own unsuccessful threat is the one who cannot threaten—but only through the paradox that he is incapable of calling a sufficiently terrible penalty on his own head.

Note that the "hurt-more" comparison in this case refers not to whether Row or Column would be hurt more by what Row threatens but to whether Row would be hurt more by having to fulfil his own threat than Column would be hurt if, instead, Column had made *his* threat. Actually, in the particular payoff matrix shown, Row's *successful* threat is one that would hurt him *more*

[19] If a player, Column, for example, cannot force first move on Row in a mechanical sense, he can do so in a "legal" sense by threatening to choose I unless Row *promises* to choose ii. Full analysis in this case requires attention to the penalties on promises as well as on threats. Since the physical and institutional arrangements for promises (i.e., for commitments to the second party) are generally of a quite different nature from those for unilateral commitments (i.e., commitments that the second player cannot himself dissolve), available penalties could differ drastically as between threats and promises—just as, in general, they would differ as between the first and second players. The particular payoffs shown in Fig. 4 would require penalties of at least 1 on a promise by Column or by Row. Note that in the case of a promise extracted by a threat, it is an advantage to the threatener to be able to invoke penalty and a disadvantage to the victim to be able to invoke penalty on his own breach of contract, ie., to be able to comply.

in the fulfilment than it would hurt Column, while Column's potential *unsuccessful* threat would hurt him *less* to fulfil than it would hurt Row.

Another loosening of the threat concept is to alter our assumption of rationality. Suppose there is some probability Pr for player R, and some probability Pc for player C, that he will make a mistake or an irrational move, or that he will act in an unanticipated way because the other player is mistaken about the first player's payoffs. This yields us a game in which the possible gains and losses in committing one's self to a threat must take into account the possibility that a fully committed threat will not be heeded. If, then, the potential loss that will ensue from having to carry out the threat is greater for one player than for another, there could be symmetrical circumstances—the P's being equal and the threat penalties equal for the two players—in which one player may find it advantageous to make the threat and the other player not, considering the possibility of "error." (A somewhat similar calculation may be involved if both players have opportunities for threats and there is danger of simultaneous commitment through the failure of one to observe the other's commitment and to stop in time to save both.)

This modification in the threat concept—in the rationality postulate that underlies it—goes somewhat in the direction of the "hurt-more" criterion. On the whole, though, game theory adds more insight into the strategy of bargaining by emphasizing the striking truth that the threat does *not* depend on the threatener's having less to suffer than the threatened party if the threat had to be carried out rather than by exaggerating the possible truth contained in the intuitive first impression. Threats of war, of price war, of damage suit; threats to make a "scene"; most of the threats of organized society to prosecute crimes and misdemeanors; and the concepts of extortion and deterrence generally cannot be understood except by denying the utility-comparison criterion. It is indeed the asymmetries in the threat situation, as between the two players, that make threats a rich subject for study; but the relevant asymmetries include those in the communication system, in the enforcibility of threats and of promises, in the speed of commitment, in the rationality of expected responses, and, finally (in some cases), in the relative-damage criterion.

PROMISES

Enforcible promises cannot be taken for granted. Agreements must be in enforcible terms and involve enforcible types of behavior. Enforcement depends on at least two things—some authority somewhere to punish or coerce and an ability to discern whether punishment or coercion is called for. The postwar discussions of disarmament proposals and inspection schemes indicate how difficult it may be, even if both sides should desperately desire to reach an enforcible agreement or find a persuasive means of enforcement. The problem is compounded when neither party trusts the other and each recognizes that neither trusts the other and that neither can therefore anticipate the other's compliance. Many of the technical problems of arms inspection would disappear if there were some earthly means of making enforcible promises or if the nations of the world all rendered unquestioned allegiance to some unearthly authority. But since non-compliance may be undetectable, promises of compliance could not be enforced even if punishment could be guaranteed. The problem is doubled by the fact that punishment cannot be guaranteed, except such punishment as can unilaterally be meted out by the other party in its act of denouncing the original agreement.

Furthermore, some seemingly desirable agreements must be left out for being undefinable operationally; agreements not to discriminate against each other will work only if defined in objective terms capable of objective supervision.

Promises are generally thought of as bilateral (contractual) commitments, given against a quid pro quo that is often a promise in return. But there is incentive for a unilateral promise when it provides inducement to the other player to make a choice in the mutual interest. In the left-hand matrix of Figure 5, if choices are to be simultaneous, only a *pair* of promises can be effective; in the right-hand matrix, Row's promise brings its own reward: Column can safely choose II, yielding superior outcomes for both players. (If, in the left-hand matrix, moves are in turn, the player who chooses

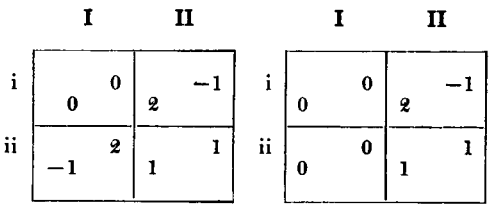

FIG. 5

second must have the power to promise. If the players are themselves to agree on the order of moves and only one of the two can issue promises, they can agree that the other one move first. These promises, in contrast to those for the right-hand matrix, must be conditional on the second player's performance. A unilateral unconditional promise does the trick on the right-hand side but not on the left with moves in turn.) The witness to a crime has a motive for unilateral promise if the criminal would kill to keep him from squealing.[20] A nation known to be on the threshold of an absolutely potent surprise-attack weapon may have reason to foreswear it unilaterally—if there is any possible way to do so—in order to forestall a desperate last-minute attempt by an enemy to strike first while he still has a chance (34).

The exact definition of a promise—e.g., in distinction to a threat—is not obvious. It might seem that a promise is a commitment (conditional or unconditional) that the second party welcomes, one that is mutually advantageous, as in both the games shown in Figure 5. But Figure 6 shows a situation in which Row must couple a threat and a promise; he threatens ii against I and promises i in the event of II. The promise insures Column a payoff of 4 rather than zero, once he has made a choice of II, and in that sense

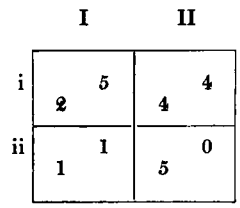

FIG. 6

it is favorable to him; it does so at a cost of 1 unit to Row. But if Row could not make the promise, Column would win 5; he would because the threat would be ineffectual without the promise, and the threat would not be incurred. A threat of ii against I by itself is no good; it cannot force Column to choose II, since a choice of II leaves him with an outcome at ii,II, zero instead of 1. Row's threat can work only if the promise

[20] This notion is celebrated in "Wet Saturday," by John Collier (5, pp. 171–78), recently reproduced by Alfred Hitchcock on TV. An inadvertent eavesdropper on a murder is ordered at gunpoint to seal his lips by leaving his own fingerprints and other incriminating evidence, so that if the body is found he will be charged with the murder. He should have insisted, however, on fabricating the evidence so as to share the guilt with the actual murderer; as it was, he got badly cheated.

goes with it; the net effect of the promise is to make the threat work, yielding Column 4 instead of 5, gaining 5 rather than 2 for Row. (One cannot force spies, conspirators, or carriers of social diseases to reveal themselves solely by the *threat* of a relentless pursuit that spares no cost; one must also promise immunity to those that come forward.)[21]

A better definition, perhaps, would make the promise a commitment that is controlled by the second party, i.e., a commitment that the second party can enforce or release as he chooses. But timing is important here. The promise just discussed will work *after* the threat is fully committed; but if the victim of the promise (Column) can renounce the promise in advance, so that Row knows that Column expects zero if he chooses II, the threat itself is deterred. And if the threat and promise are contrived in such a way as to be "legally" inseparable or if they are accomplished by some irreversible maneuver, the definition becomes obscured. (In fact, the definition breaks down whenever the equivalent of a promise is obtained by some irrevocable act rather than by a "legal" commitment.)

Actually, whenever the alternative choices are more than two, threat and promise are likely to be mixed in any "reaction pattern" that one presents to the other. So it is probably best to consider the threat and the promise to be names for different aspects of the same tactic of selective and conditional self-commitment, which in certain simple instances can be identified in terms of the second party's interest.

Enforcement schemes. Agreements are unenforcible if no outside authority exists to enforce them or if non-compliance would be inherently undetectable. The problem arises, then, of finding forms of agreement, or terms to agree on, that provide no incentive to cheat or that make non-compliance automatically visible or that incur the penalties on which the possibility of enforcement rests. While the possibility of "trust" between two partners need not be ruled out, it should also not be taken for granted; and even trust itself can usefully be studied in game-theoretic terms. Trust is often achieved simply by the continuity of the relationship between parties and the recognition by each that what he might gain by cheating in a given instance is outweighed by the value of the tradition of trust that makes possible a long sequence of future agreement. By the same token, "trust" may be achieved for a single discontinuous instance, if it can be divided into a succession of increments.

There are, however, particular game situations that lend themselves to enforcible agreement. One is an agreement that depends on some kind of co-ordination or complementarity. If two people have disagreed on where to meet for dinner; if two criminal accomplices have disagreed on what joint alibi to give; or if members of a business firm or football team have disputed about what prices they will quote or what tactic they will follow, they nevertheless have an overriding interest in the ultimate consistency of their actions. Once agreement is formally reached, it constitutes the only possible focal point for the necessary subsequent tacit collaboration; no one has a unilateral preference now to do anything but what he is expected to do. In the absence of any other means of enforcement, then, parties might be well advised to try to find agreements that enjoy this property of interdependent expectations, even to the extent of importing in their agreement certain ele-

[21] Somewhat related is the grant of immunity that strips a reticent witness of protective danger of self-incrimination, and so opens him to the ordinary sanction of contempt proceedings.

ments whose sole purpose is to create severe jeopardy for non-co-ordination. Tearing the treasure map in half or letting one partner carry the gun and the other the ammunition is a familiar example.

The institution of *hostages* is an ancient technique that deserves to be studied by game theory, as does the practice of drinking wine from the same glass or of holding gang meetings in places so public that neither side could escape if it subjected the other to a massacre. The reported use of only drug addicts as agents or employees in a narcotics ring is a fairly straightforward example of a unilateral hostage.

Perhaps a sufficient interchange of populations between nations that hate each other or an agreement to move the governing agencies of both countries to a single island where they would occupy alternate blocks of the city could be resorted to if both sides became sufficiently desperate to avoid mutual destruction. A principal drawback to the exchange of hostages, on the assumption of rational behavior, is the inherent unknowability of each other's value system adverted to earlier. The king who sends his daughter as a hostage to his enemy's court may be incapable of assuaging his enemy's fears that he really dislikes the girl. We could probably guarantee the Russians against an American surprise attack by having the equivalent of "junior year abroad" at the kindergarten level: if every American five-year-old went to kindergarten in Russia —in American establishments constructed for the purpose, designed solely for "hostage" purposes and not for cultural interchange— and if each year's incoming group arrived before the graduating class left, there would not seem to be the slightest chance that America would ever initiate atomic destruction in Russia. We cannot be quite sure that the Russians would be quite sure of this. Nor can we be quite sure that a reciprocal program would be as much of a deterrent to the Russian government; unfortunately, even if the Russian government were bound by the fear of harming Russian children, it seems nearly impossible for it to persuade us so. Still, in many surprise-attack situations a unilateral promise is better than none; and the idea of hostages may be worth considering, even when symmetrical exchanges do not seem available.[22]

Actually, the hostage idea is logically identical with the notion that a disarmament agreement between the major powers might be more efficacious (and probably more subject to technical control) if it related to *defensive* weapons and structures. To eschew defense is, in effect, to make hostages of your entire population without bothering to put them physically into the other's possession. Thus we can put our children at the mercy of the Russians and receive similar power over Russian children not only by physically trading them, with enormous discomfort and breach of constitutional rights, but also by simply agreeing to leave them so unprotected that the other can do them as much damage where they are as if he had them in his grasp. Thus the "balance of terror" that is so often adverted to is—if, in fact, it exists and is stable—equivalent to a total exchange of all conceivable hostages. (The analogy requires that the balance be

[22] The precise definition of hostages is a little difficult. They seem to be as pertinent to threats as to promises: the four American divisions that were stationed in Europe principally to demonstrate that America could not avoid becoming engaged in a European conflict can probably be viewed as hostages; if they cannot, their wives and children can, and perhaps their wives and children have been a more persuasive commitment of "trip wire" than the troops themselves. As a general rule, invaders may have to avoid the peak tourist season in countries they covet, to avoid provoking the countries that have yielded inadvertent hostages.

stable, i.e., that neither side be able, by surprise attack, to destroy the other's power to strike back, but just able to inflict a surfeit of civilian agony.)

Denial of enforcement. Enforcement of promises is also relevant to the influence of a third party that wishes to make an efficient outcome more difficult for the other two players. A potent means of banning illegal activities has often been the outlawing of them, so that contracts became unenforcible. Failure to enforce gambling contracts or contracts in restraint of trade or contracts for the delivery of liquor during prohibition has always been part of the process of discouraging the activities themselves. Sometimes, of course, prohibition of this sort delivers enormous power into the hands of anyone who can enforce contracts or make enforcible promises.[23] The denial of copyright liquor labels during prohibition meant that only the bigger gangs could guarantee the quality of their liquor and hence assisted them in developing monopoly control of the business. By the same token, laws to protect brands and labels can perhaps be viewed as devices that facilitate business based on unwritten contracts.

RELINQUISHING THE INITIATIVE

What makes the threat or ordinary commitment a difficult tactic to employ and an interesting one to study is the problem of finding a means to commitment, the available "penalty" to invoke against one's own non-performance. There is consequently a related set of tactics that consists of maneuvering one's self into a position in which one no longer has any effective choice over how he shall behave or respond. The purpose of these tactics is to get rid of an embarrassing initiative, making the outcome depend solely on the other party's choice.

This is the kind of tactic that Secretary of State John Foster Dulles (8) was looking for in the following passage:

> In the future it may thus be feasible to place less reliance upon deterrence of vast retaliatory power. . . . Thus, in contrast to the 1950 decade, it may be that by the 1960 decade the nations which are around the Sino-Soviet perimeter can possess an effective defense against full-scale conventional attack and thus confront any aggressor with the choice between failing or himself initiating nuclear war against the defending country. Thus the tables may be turned, in the sense that instead of those who are non-aggressive having to rely upon all-out nuclear retaliatory power for their protection, would-be aggressors would be unable to count on a successful conventional aggression, but must themselves weigh the consequence of invoking nuclear war.[24]

The distinction between the type of deterrence he imputes to the 1950's and the type he imputes to the 1960's differs in the matter of who has to make that final decision; and the difference is important because the United States cannot find, or bring itself to trust, a persuasive means of commitment to the threat of massive retaliation against certain types of aggression.

There was a time, shortly after the first atomic bomb was exploded, when there was some journalistic speculation about whether the earth's atmosphere had a limited tolerance to nuclear fission; the idea was bruited

---

[23] It has been argued that an important function of the racketeer is sometimes to help enforce agreements that are beyond the law. Price-cutting in the Chicago garment trade was punishable by explosion—the fee for the explosion being paid by the price-fixing organization—according to R. L. Duffus (7).

[24] Very similar language is used by Dean Acheson (1, pp. 87–88) in discussing the role of a sizable defense force in Europe: by requiring of the enemy a major attack, rather than a small one, it makes him believe that retaliation would ensue, because "he would be making the decision for us. . . . A defense in Europe of this magnitude will pass the decision to risk everything from the defense to the offense."

about that a mighty chain reaction might destroy the earth's atmosphere when some critical number of bombs had already been exploded. Someone proposed that, if this were true and if we could calculate with accuracy that critical level of tolerance, we might neutralize atomic weapons for all time by a deliberate program of openly and dramatically exploding $n - 1$ bombs.

This tactic of shifting responsibility to the other player was nicely accomplished by Lieutenant Colonel (then Major) Stevenson B. Canyon, U.S.A.F., in using his aircraft to protect a surface vessel about to be captured by Communist surface forces in his comic strip. Unwilling and unauthorized to initiate hostilities and knowing that no threat to do so would be credited, he directed his planes to jettison gasoline in a burning ring about the aggressor forces, leaving to them the last clear chance of reversing their engines. He could neither drop gasoline on the enemy ships nor threaten to; so he dropped the initiative instead.

The same tactic is involved in those dramatic forms of "passive resistance" that might be better called "active non-resistance." According to *The New York Times* (29), "Striking railway workers sat down on the tracks at more than 300 stations in Japan today, halting 48 passenger and 144 freight trains."[25]

A more dramatic instance, also Japanese, was reported in the same paper (28): "A public debate is being held here this week on whether to send a 'suicide sit-down fleet' to the forbidden waters around Christmas Island, the site of the forthcoming British hydrogen bomb experiment. . . . The first object of the expedition would be to prevent the British blast."

IDENTIFICATION

An important characteristic of any game is how much each side knows about the other's value system; but a similar information problem arises with respect to sheer identification. The bank employee who would like to rob the bank if he could only find an outside collaborator and the bank robber who would like to rob the bank if only he could find an inside accomplice may find it difficult to collaborate because they are unable to identify each other, there being severe penalties in the event that either should declare his intentions to someone who proved not to have identical interests. The boy who is afraid to ask a girl for a date because she might rebuff him is in a similar position. Similarly, the kidnaper cannot operate properly if he cannot tell the rich from the poor in advance; and the antisegregation minority in the South may never know whether it is large or small because of the penalties on declaration.

Identification, like communication, is not necessarily reciprocal; and the act of self-identification may sometimes be reversible and sometimes not. One may achieve more identification than he bargained for, once he declares his interest in an object. A nice example occurs in Shakespeare's *Measure for Measure*. The acting Duke has a prisoner whom he proposes to kill. He could torture him, but he has no incentive to. The victim has a sister, who arrives to plead for

[25] The appropriate countertactic seems to be the following: The engineer sets the throttle for slow forward speed, conspicuously climbs down from his cab and jumps off the moving train, walks through the station and jumps back on his engine when it catches up with him. The weakness of his position while he is driving the train is that he can stop it more quickly than his adversaries can get off the tracks, particularly if they have arranged to crowd themselves so that they could not vacate the track quickly. They can forestall his countertactic by locking themselves to the tracks and throwing away the key—if they can persuasively inform the engineer of this before he has relinquished his own control of the engine.

his life. The Duke, finding the sister attractive, proposes a dishonorable bargain; the sister declines. The Duke then threatens to torture the brother unless the sister submits. At this point the game has been expanded simply by the establishment of identity and of a line of communication. The Duke's only interest in torturing the brother is in what he may gain by making a threat to do so; once there is somebody available to whom the threat can profitably be communicated, the possibility of torture has value for the Duke—not the torture itself, but the threatening of it. The sister has gotten negative value out of her trip; having identified her interest and made herself available to receive the threatening message, she has been forced to suffer what she would not have had to suffer if she had never made her identity known or if she could have disappeared into the crowd before the threat was made.

A nice identification game was uncovered in a New York suburb a few years ago. Certain motorists carried identity cards which identified them to policemen as members in a club; if the motorist with a membership card was arrested, he simply showed the card to the policeman and paid a bribe. The role of these cards was to identify the motorist as a person who, if the bribe was received, would keep quiet. It identified the motorist as a man whose promise was enforcible. But the card identifies the motorist only *after* he has been arrested; if the police could identify card-carrying motorists by looking at them, they could concentrate their arrests on card-carrying drivers, threatening a ticket unless payment were received. The card is contingent identification, at the option of the motorist. A similar situation —pertinent to the discussion of promises as well as to identification—is described by Sutherland (36, p. 126): "Most coppers are more or less fair in their dealings with thieves simply because it pays them to be so. They will extend favors even after a pinch which they would not extend to nonprofessionals whom they lock up. They realize that it is safe to do this and that high officials will not be informed, as might be the case if favors were extended to amateurs."

Identification is also relevant to an important economic fact that tends to be ignored in the conventional economics of production and exchange, namely, the enormous potential for destruction that is available and that is relevant because of the extortionate threats that could be supported by it. The ordinary healthy high-school graduate, of slightly below average intelligence, has to work fairly hard to produce more than $3,000 or $4,000 of value per year; but he could destroy a hundred times that much if he set his mind to it, according to the writer's hasty calculations. Given an institutional arrangement in which he could generously abstain from destruction in return for a mere fraction of the value that he might have destroyed, the boy clearly has a calling as an extortionist rather than as a mechanic or clerk. It is fortunate that extortion usually depends on self-identification and overt communication by the extortionist himself.

The importance of self-identification is attested by the significance attached to the doctrine that an accused person should be permitted to know and to confront his accuser. It is also reflected in secret testimony before a Grand Jury, in cases where identifiable witnesses might be intimidated by potential defendants, and in efforts to keep secret the identity of eyewitnesses to a crime until the criminal is apprehended. (The strategy of law and of law enforcement and criminal deterrence is a rich field for the application of game theory.)

DELEGATION

Another "move" that is sometimes available is the delegation of part or all of one's interest, or part or all of one's initiative for decision, to some agent who becomes (or perhaps already is) another player in the game. Insurance schemes permit the sharing of interests; the insurance company has a different incentive structure from the insured party and may be better able to make threats or resist them for that reason. Requiring several signatures on a check accomplishes a similar purpose. The use of a professional collecting agency by a business firm for the collection of debts is a means of achieving unilateral rather than bilateral communication with its debtors and of being therefore unavailable to hear pleas or threats from the debtors. Providing ammunition to South Korean troops or giving them access to prisoner-of-war camps so that they can unilaterally release prisoners is a tactical means of relinquishing an embarrassing power of decision—embarrassing because it subjects one to coercive or deterrent threats or leaves one the capacity to back out of his own threat, hence the incapacity to make the threat persuasive.

The mutual-defense agreement with the Nationalist government of China is probably to be viewed partly as a means of shifting the decision for response to someone whose resolution would be less doubtful; and more recently the proposal to put nuclear weapons in the hands of European governments has been explicitly argued on grounds that it would enhance deterrence by giving the visible power to retaliate to countries that might in certain contingencies be thought less irresolute than the United States.

The use of thugs and sadists for the collection of extortion or the guarding of prisoners, or the conspicuous delegation of authority to a military commander of known motivation, exemplifies a common means of making credible a response pattern that the original source of decision might have been thought to shrink from or to find profitless, once the threat had failed. (Just as it would be rational for a rational player to destroy his own rationality in certain game situations, either to deter a threat that might be made against him and that would be premised on his rationality or to make credible a threat that he could not otherwise commit himself to, it may also be rational for a player to select irrational partners or agents.)

In the matrix in Figure 7—disregarding

	I	II
i	3 (2) 5	2 (1) 0
ii	4 (1) 0	5 (0) 1

FIG. 7

the numbers in parentheses—if Row has second move, he loses in the lower right-hand corner, Column gaining his own preferred outcome. If a third party without power of decision is scheduled to receive, as a by-product, the payoff in parentheses, Row can win if some means is available for irreversibly surrendering his move to the third player. The payoffs of the latter are such that with second move he wins in the upper left-hand corner, leaving the original Row-player a payoff of 5 as a by-product. (If the third party's rewards had to be financed by Row, whose own payoffs were correspondingly reduced, it would still be worth his while to make an irrevocable assignment of portions of his various payoffs to the third player, together with assignment of the decision; with the figures shown, he would

still carry away a net value of 3 in the upper left-hand corner, in contrast to 1 in the lower right.)

MEDIATION

The role of mediator is another element for analysis in game theory. A mediator, whether imposed on the game by its original rules or adopted by the players to facilitate an efficient outcome, is probably best viewed as an element in the communication arrangements or as a third player with a payoff structure of his own who is given an influential role through his control over communication. But a mediator can do more than simply constrain communications—putting limits on the order of offers, counteroffers, etc.—since he can invent contextual material of his own and make potent suggestions. That is, he can influence the other player's expectations on his own initiative, in a manner that both parties cannot help mutually recognizing. When there is no apparent focal point for agreement, he can create one by his power to make a dramatic suggestion. The bystander who jumps into an intersection and begins to direct traffic at an impromptu traffic jam is conceded the power to discriminate among cars by being able to offer a sufficient increase in efficiency to benefit even the cars most discriminated against; his directions have only the power of suggestion, but co-ordination requires the common acceptance of some source of suggestion. Similarly, the participants of a square dance may all be thoroughly dissatisfied with the particular dances being called, but as long as the caller has the microphone, nobody can dance anything else. The white line down the center of the road is a mediator, and very likely it can err substantially toward one side or the other before the disadvantaged side finds advantage in denying its authority.

Mediators can also be a means by which rational players can put aside some of their rational faculties. A mediator can consummate certain communications while blocking off certain facilities for memory. (In this regard he serves a function that can be reproduced by a computing machine.) He can, for example, compare two parties' offers to each other, declaring whether or not the offers are compatible without revealing the actual offers. He is a scanning device that can suppress part of the information put into it. He makes possible certain limited comparisons that are beyond the mental powers of the participants, since no player can persuasively commit himself to forget something.

The problem of persuasively denying one's self the knowledge that one receives by the left hand, while actively seeking it with the right hand, is nicely illustrated by the efforts of parts of governments to obtain accurate data on incomes for the purpose of statistical programs, while another part of the government is seeking the same data in order to impose taxes or to prosecute evasion. Governments have found it important to seek ways of guaranteeing that the statistical agency will deny the information it receives to the taxing agency, in order to receive the information in the first place. An analogous case of relying on an explicit mediator is that of companies that turn trade secrets over to a statistical bureau that is committed to destroy the individual data after computing the sums and averages that it will make public for the benefit of the contributing companies, or of public opinion services that suppress potentially embarrassing individual data on political or sexual practices, publishing only the aggregates. The use of mediators to forestall identification seems to be a common tactic when a buyer of large resources thinks a painting,

or a right-of-way can be bought cheap if the owner is unaware who it is that is interested.

Mediators may be converted into arbitrators by the irrevocable surrender of authority to him by the players. But arbitration agreements have to be made enforcible by the players' deliberately incurring jeopardy, providing the referee with the power to punish or surrendering to him something complementary to their own value systems. In turn, they must be able to trust him or to extract an enforcible promise from him. But in any case he increases the totality of means for enforcing promises: two people who do not trust each other may find a third person that they both trust, and let him hold the stakes.[26]

COMMUNICATION AND ITS DESTRUCTION

Many interesting game tactics and game situations depend on the structure of communication, particularly asymmetries in communication and unilateral options to initiate communication or to destroy it. Threats are no good if they cannot be communicated to the persons for whom they are intended; extortion requires a means of conveying the alternatives to the intended victim. Even the threat, "Stop crying or I'll give you something to cry about," is ineffectual if the child is already crying too loud to hear it. (It sometimes appears that children know this.) A witness cannot be intimidated into giving false testimony if he is in custody that prevents his getting instructions on what to say, even though he might infer the sanction of the threat itself.

When the outcome depends on co-ordination, the timely destruction of communication may be a winning tactic. When a man and his wife are arguing by telephone over where to meet for dinner, the argument is won by the wife if she simply announces where she is going and hangs up. And the status quo is often preserved by a person who evades discussion of alternatives, even to the extent of simply turning off his hearing aid.

As discussed in the earlier part of this paper, mob action often depends on communication in a way that makes it possible for the authorities to obstruct mob action by forbidding groups of three or more to congregate. But mobs can themselves intimidate the authorities if they are able to identify them and to communicate with them. Even a tacit threat of subsequent ostracism or violence may be communicated from a riotous mob to the local police, if the police are known to them and are persons who have to reside among them when the occasion is over. In that case the use of outsiders may forestall the mob's intimidating threats against the authorities, partly by reducing the subsequent occasion for carrying out the threat but partly also through the difficulty of tacit communication between mob and police. Federal troops in Little Rock may have enjoyed some immunity to intimidation just by being outside the tacit communication structure of the local populace and being patently less conversant with the local value system than were the local police. State troops were dramatically successful in quelling the Detroit race riot of 1943, when the local police were ineffectual. The use of Moors, Sikhs; and other foreign-language

---

[26] I have been told that in countries where no strong tradition of business morality exists, a few partners or directors for a business may deliberately be chosen from another culture where simple honesty and fairness are considered to be common traits or where a reputation for them is considered of much higher value.

troops against local uprising may owe some of its success to their poor capacity to receive the threats and promises that the enemies or victims might otherwise seek to convey. Even the isolation of officers from enlisted men in military service may tend to make officers less capable of receiving and perceiving threats, hence less capable of being effectively threatened, and thus deterring intimidating threats themselves.

It is important, of course, whether or not the threatener knows that his threat cannot be received; for if he thinks it can, and it cannot, he may make the threat and fail in his objective, being obliged to carry out his threat to the subsequent disadvantage of both himself and the one threatened. So the soldiers in quelling the riot should not only be strangers and not only keep moving sufficiently to avoid "acquaintance" with particular portions of the mob; they should behave with an impassivity to demonstrate that no messages are getting through. They must catch no one's eye; they must not blush at the jeers; they must act as if they cannot tell one rioter from another, even if one has been making himself conspicuous. Figuratively, if not literally, they should wear masks; even the uniform contributes to the suppression of identification and so itself makes reciprocal communication difficult.

Conveyance of evidence. "Communication" refers to more than the transmission of messages. To communicate a threat, one has to communicate the commitment that goes with it, and similarly with a promise; and to communicate a commitment requires more than communication of words. One has to communicate *evidence* that the commitment exists; this may mean that one can communicate a threat only if he can make the other person see something with his own eyes or if he can find a device to authenticate certain allegations. One can send a signed check by mail, but one cannot demonstrate over the telephone that a check bears an authentic signature; one may show that he has a loaded gun but not prove it by simply saying so. From a game-theory point of view, the Paris *pneumatique* differs from an ordinary telegraph system, and television differs from radio. (One role of a mediator may be to authenticate the statements that the players make to each other; for example, a code system for identification might make it possible for people to transmit funds orally by telephone, the recipient being assured by the bank's code response that it is in fact the bank at the other end of the line assuring him that the payer has been identified by code and that the transaction is complete.) The importance and the difficulty of communicating evidence is exemplified by President Eisenhower's "open-skies" proposal and other suggested devices for dealing with the instability that may be caused by the reciprocal fear of surprise attack. Leo Szilard (37) has even pointed to the paradox that one might wish to confer immunity on foreign spies rather than subject them to prosecution, since they may be the only means by which the enemy can obtain persuasive evidence of the important truth that we are making no preparations for embarking on a surprise attack.

It is interesting to observe that political democracy itself depends on a game structure in which the communication of evidence is impossible. What is the secret ballot but a device to rob the voter of his power to sell his vote? It is not alone the secrecy, but the *mandatory* secrecy, that robs him of his power. He not only *may* vote in secret, but he *must* if the system is to work. He must be denied any means of proving which way he voted. And what he is robbed of is not just an asset that he

might sell; he is stripped of his power to be intimidated. He is made impotent to meet the demands of blackmail. There may be no limit to violence that he can be threatened with if he is truly free to bargain away his vote, since the threatened violence is not carried out anyway if it is frightening enough to persuade him. But when the voter is powerless to prove that he complied with the threat, both he and those who would threaten him know that any punishment would be unrelated to the way he actually voted. And the threat, being useless, goes idle.

An interesting case of tacit and asymmetrical communication is that of a motorist in a busy intersection who knows that a policeman is directing traffic. If the motorist sees, and evidently sees, the policeman's directions and ignores them, he is insubordinate; and the policeman has both an incentive and an obligation to give the man a ticket. If the motorist avoids looking at the policeman, cannot see the directions, and ignores the directions that he does not see, taking a right of way that he does not deserve, he may be considered only stupid by the policeman, who has little incentive and no obligation to give the man a ticket. Alternatively, if it is evident that the driver knew what the instructions were and disobeyed them, it is to the policeman's advantage not to have seen the driver, otherwise he is obliged, for the reputation of the corps, to abandon his pressing business and hail the driver down to give him a ticket. Children are skilled at avoiding the receipt of a warning glance from a parent, knowing that if they perceive it the parent is obliged to punish non-compliance; adults are equally skilled at not requesting the permission they suspect would be denied, knowing that explicit denial is a sterner sanction, obliging the denying authorities to take cognizance of the transgression.[27]

The efficacy of the communication structure can depend on the kinds of rationality that are imputed to the players. This is illustrated by the game situation known as "having a bear by the tail." The minimum requirement for an efficient outcome is that the bear be able to incur an enforcible promise and that he be able to transmit credible evidence that he is committed, either by a penalty incurred or by a maneuver that destroys his power not to comply (like extracting his own teeth and claws). But if the bear is of limited rationality, having a capacity for making rational and consistent choices among the alternatives that he perceives but lacking the capacity to solve games—i.e., lacking the capacity to determine introspectively the choices that a partner would make—the communication system must make it possible for him to receive a message from his partner. The partner must then formulate

[27] What might be called the "legal status" of communication is nicely developed by Goffman (11, pp. 224, 226): "Tact in regard to face-work often relies for its operation on a tacit agreement to do business through the language of hint—the language of innuendo, ambiguities, well placed pauses, carefully worded jokes, and so on. The rule regarding this unofficial kind of communication is that the sender ought not to act as if he had officially conveyed the message he has hinted at, while the recipients have the right and the obligation to act as if they have not officially received the message contained in the hint. Hinted communication, then, is deniable communication." He refers to the "unratified" participation that can occur in spoken interaction: "A person may overhear others unbeknown to them; he can overhear them when they know this to be the case and when they choose either to act as if he were not overhearing them or to signal to him informally that they know he is overhearing them." He points out that the obligation to respond, for example, to an insulting remark that one has inadvertently overheard may depend on whether the overhearing has acquired "ratification."

the proposition (choice) for the bear and communicate it to him, in order that the bear may then respond by accepting the promise (now that he sees what the "solution" is) and transmitting authoritative evidence back to his own partner.

INCORPORATION OF MOVES IN A GAME MATRIX

One is led to suppose that, if a game has potential moves like threats, commitments, and promises that are susceptible of formal analysis, it must be possible to represent such moves in the traditional form of strategy choices, with the payoff matrix of the original game expanded to allow for the choices among these various moves.

The first point to observe is that a commitment, a promise, or a threat can usually be characterized in a fashion equivalent to the following: to make one of these moves, a player selectively reduces—visibly and irreversibly—some of *his own* payoffs in the matrix. This is what the move amounts to. We could also say that one openly selects a strategy in advance for responding to the other's choice; but more than selection is required. The player must invoke penalty on his own failure to pursue subsequently the particular strategy of response that he has selected beforehand. And to invoke a penalty on failure to follow a strategy is mathematically equivalent to subtracting the amount of the penalty from one's own payoffs in all cells that do not correspond to the strategy so selected.[28]

Specifically, in Figure 8, *A*, Row would commit himself to ii by subtracting from his own payoffs in the first row sufficiently large quantities—5 in the example shown—to make ii a dominant strategy, i.e., a strategy that he would follow no matter which column the other player selects. The result would be the modified matrix shown in Figure 8, *B*. (Committing himself to i

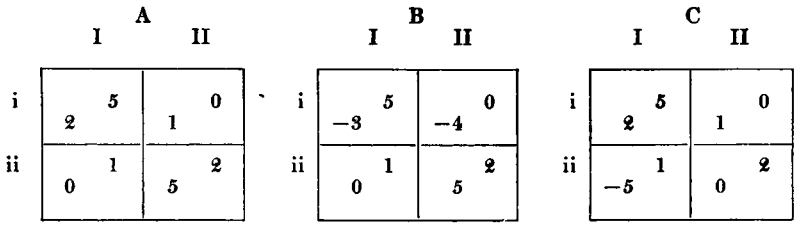

FIG. 8

[28] Threats, promises, and unconditional commitments have already been illustrated; a more general "reaction function" is illustrated in the accompanying matrix. If Row can attach adequate penalties to his own selection of any cells other than those starred, he leaves Column a simple maximization problem which Column solves by choosing his third strategy. Row has "won" almost his favorite cell; specifically, he has secured for himself the most favorable cell among those that leave Column no lower than his "minimax" value. This is the generalization of the tactic that, for simple two-way or three-way choices, can be identified as a "commitment," "threat," "promise," or combination of them.

	I	II	III	IV	V
i	6 1	10 11	2 10	9 2	7 10
ii	9 8	*4 12	0 25	1 20	15 3
iii	20 9	15 2	*6 16	*1 18	17 14
iv	*2 6	10 8	7 7	4 5	*3 20

with penalty of 5 would yield the matrix in Fig. 8, *C*.) Can we now build up a larger matrix that represents not only the actual choices of rows and columns in the original game, such as those in Figure 8, *A*, but also the strategies of *commit, threaten, promise,* etc.? Certainly, once we have specified what moves are available and the order in which they are to be taken. Take the simple game in which Row has the power to commit himself visibly in advance, and Column has first move in the *original* game, that is, chooses his column before Row makes his *final* choice of row.

Originally Row, having second move, had four strategies available. He could pick i no matter what; he could pick ii no matter what; he could play i to column I and ii to column II; or he could play ii to column I and i to column II. Including the possibility of commitment, he now has *first* the choice of committing himself; and to each of these first choices he can attach any one of the four strategies just mentioned for his final move. For example, he can commit himself to ii and play ii no matter what; he can commit himself to ii and play i no matter what; he can commit himself to ii and play

		I	II	III	IV	V	VI	VII	VIII
		0-I 1-I 2-I	0- I 1- I 2-II	0- I 1-II 2- I	0- I 1-II 2-II	0-II 1- I 2- I	0-II 1- I 2-II	0-II 1-II 2- I	0-II 1-II 2-II
i	0; I-i, II-i	5 2	5 2	5 2	5 2	0 1	0 1	0 1	0 1
ii	0; I-ii, II-ii	1 0	1 0	1 0	1 0	2 5	2 5	2 5	2 5
iii	0; I-i; II-ii	5 2	5 2	5 2	5 2	2 5	2 5	2 5	2 5
iv	0; I-ii, II-i	1 0	1 0	1 0	1 0	0 1	0 1	0 1	0 1
v	1; I-i, II-i	5 2	5 2	0 1	0 1	5 2	5 2	0 1	0 1
vi	1; I-ii, II-ii	1 −5	1 −5	2 0	2 0	1 −5	1 −5	2 0	2 0
vii	1; I-i, II-ii	5 2	5 2	2 0	2 0	5 2	5 2	2 0	2 0
viii	1; I-ii, II-i	1 −5	1 −5	0 1	0 1	1 −5	1 −5	0 1	0 1
ix	2; I-i, II-i	5 −3	0 −4	5 −3	0 −4	5 −3	0 −4	5 −3	0 −4
x	2; I-ii, II-ii	1 0	*2 5	1 0	*2 5	1 0	*2 5	1 0	*2 5
xi	2; I-i, II-ii	5 −3	2 5	5 −3	2 5	5 −3	2 5	5 −3	2 5
xii	2; I-ii, II-i	1 0	0 −4	1 0	0 −4	1 0	0 −4	1 0	0 −4

FIG. 9

i to column I, ii to column II; or he can commit himself to ii and play ii to column I, i to column II. Altogether, he has twelve possible strategy combinations.

Column has eight possible strategy combinations: for each of three contingencies he has either of two moves, the moves being I and II, the contingencies being Row's commitment to i, Row's commitment to ii, and Row's non-commitment.

If we put these strategies into matrix form, we get Figure 9. The 12×8 matrix of Figure 9 represents the tacit ("non-cooperative") game that corresponds to the players' private decisions on *how to play* the original game. The eight possible strategies available to Column, for example, can be thought of as the eight possible distinct *sets of complete instructions* that he might give an agent who would then play the original game for him—that is, play the game at which he chooses one of two columns, depending on whether and how Row committed himself first. There is no loss to either player in being supposed to play this enlarged game tacitly, since what would have been each player's *adaptations* to the other's prior moves is now fully allowed for in the specification of strategies in the enlarged version of the game; they *are* strategies of response or adaptation.

This is brought out in the labeling of Figure 9. As before, Column's choices in the original two-move game are labeled I and II; Row's choices, i and ii. Additionally, the symbol "2" will denote Row's commitment to row ii, "1" a commitment to row i, and "0" a decision not to commit himself. In the enlarged game, a single "strategy" for Column is now denoted by three pairs of symbols, such as 0-I, 1-II, 2-I, which would mean, "Choose column I if he does not commit himself, column II if he commits himself to row 1, and column I if he commits himself to row 2." For Row, a strategy consists of a decision on 0, 1, or 2, plus a pair of symbols denoting how he will react to each of Column's possible choices. For example, 1; I-i, II-i would mean, "Commit to row 1, then choose row 1 no matter what Column does." Knowing the payoffs in the original game, Figure 8, *A*, the players can identify the payoffs in the enlarged game of Figure 9. We can imagine Row and Column, instead of meeting to play the original game, sending their agents to play for them, each agent fully instructed for all contingencies (i.e., given one particular strategy for the enlarged game). To determine what instructions to give, Row and Column consider the matrix in Figure 9; in effect, they play the tacit game in that matrix, leaving to their agents just the role of messenger.

What is the "solution" of this enlarged tacit game? Or, rather, can we identify an evident solution to the original game? And, if so, how does it show up in the enlarged matrix? The original game clearly has a solution for rational players. (*A*) If Row is committed to row i, with a penalty of 5 for breaking his commitment, Column can see that row i will be chosen, no matter which column he chooses; Column chooses his preferred cell in the upper row, which is the upper left cell, i,I. And Row knows that, if he commits himself to row i, he gets the payoff in that upper-left cell, which is 2. (*B*) If, instead, Row commits himself to row ii (subtracts 5 from his payoff in row i), Column chooses II in preference to I; and Row knows he will get 5. Finally, (*C*) if Row remains uncommitted, Column knows that Row will pick the highest row payoff in the column chosen; thus if Column chooses I, Row takes i, and Column gets 5; if Column takes II, Row takes ii, and Column gets 2. Column prefers I; this leaves Row a payoff of 2; and Row can anticipate it. So Row's best outcome is to commit

himself to row ii. This is the evident "solution"; it has a payoff of [$_5$ ^{2}], and it corresponds to the strategy 2; I-ii, II-ii for Row, and to all four strategies containing 2-II for Column. (What Column would have done in contingencies 0 and 1 is of no material consequence, once Row has made his first move.) These are the starred cells in Figure 9, row x. (In effect, Row's first move is a choice of which to play among the three different two-move games, *A*, *B*, and *C*, shown in Figure 8, in which he has second move.)

How do we characterize the cells, or pairs of strategies, that represent the "solution" in Figure 9? They constitute a solution of the kind that is called a *solution in the complete weak sense* (21, pp. 106–9). It is arrived at by a process of discarding "dominated" rows and strategies. A row is dominated by another row if every payoff to Row in the dominating row is at least as good as the corresponding payoff in the dominated row and at least one payoff is better. Applying this criterion, the first row is dominated by the third, and we strike it out. (The argument might be that Row can safely eliminate the strategy represented in the first row, since the third is at least as good in every contingency and better in some.) So is the second, so is the fourth; so are all the rest except the tenth. Neither the third nor the tenth row dominates the other, so for the moment we keep them both. Comparing columns, no single column dominates another; but, having eliminated all rows but the third and tenth (arguing, perhaps, that Row would not choose them anyway), Column can make his comparison between only the third and tenth cells in the columns. Now it is apparent that the second column dominates the first, the third, the fifth, and the seventh. After striking out those columns that are dominated in the reduced set of rows, we can look again at rows iii and x. Originally, neither dominated the other; but, with the first, third, fifth, and seventh *columns* gone, the tenth row dominates the third. Striking out the third row, we are left with a single row, row x, intersected by four columns. The payoffs are the same in the four intersections, indicating that it is inconsequential which of those four strategies Column plays, as long as Row plays the tenth row. (That is, once Row has committed himself to the second row, it makes no difference what instructions Column gives his agent regarding the two contingencies that did not arise.)

This, then, is the way that a solution to the original *sequential-move game* shows up in the static ("moveless," or simultaneous-tacit-choice) game. It is a solution arrived at by discarding dominated strategies, with the criterion for domination reflecting only the undiscarded strategies at each stage. This seems to be the general form of solution in the enlarged tacit game that corresponds to a sequential-move game when the latter has a determinate solution. The discarding of rows and columns can actually be identified with the process of first calculating the rational *last move* for all possible sets of prior moves, then, knowing what last move would follow each next-to-last move, calculating the best next-to-last move for all possible sets of prior moves and so on back to the best first move of the game.

While it is instructive and intellectually satisfying to see how such tactics as threats, commitments, and promises can be absorbed into an enlarged, abstract "supergame" (game in "normal form"), it should be emphasized that we cannot learn anything about those tactics by studying games that are already in normal form. The objects of our study, namely, these tactics together with the communication and en-

forcements structures that they depend on, and the timing of moves, have all disappeared by the time the game is in normal form. What we want is a theory that systematizes the study of the various universal ingredients that make up the move-structure of games; too abstract a model will miss them.[29]

The matrix representation of a sequential game does help emphasize, however, that the formal "determinateness" of games that are resolved by tactical moves does not detract from their essential game-of-strategy character. A threat "wins" and determines an outcome only because it induces the other player to choose in one's favor. The other player retains his original freedom of choice; and his choice still depends on his anticipation of the threatener's final choice. The threatener's first choice—to threaten or not—thus depends on what he expects the threatened player to expect the threatener to do. The reciprocal-expectation character of the game remains; the threat, like the unconditional commitment or like the broader concept of reaction function when many choices of action are available, works by constraining another player's expectations through the manipulation of one's own incentives.

THE PARADOX OF STRATEGIC ADVANTAGE

It is, of course, a corollary principle that if the payoff matrix to begin with had already shown values for one of the players reduced in the same pattern as that in which he would reduce it deliberately at the winning move, he simply wins without needing to make the move overtly. This is the point that, in an earlier article by the author (30), was referred to as an abstract analogy of the principle that, in bargaining, weakness may be strength. There is probably no single principle of game theory that epitomizes so strikingly the mixed-motive game as this principle that a worsening of some or even all of the potential outcomes for a particular player and an improvement in none of them may be distinctly—even dramatically—advantageous for the player so disadvantaged. It explains why a sufficiently severe and certain penalty on the *payment* of blackmail can protect the potential victim, how the burning of bridges behind one's self may dishearten an enemy and induce his retirement, or why a lady might, in an earlier era, defy the search party by haughtily placing the sought object in her bosom.[30]

It was reported unofficially during the Korean War that when the Treasury Department blocked Communist Chinese financial assets, it also knowingly blocked some non-Communist assets as a means of

[29] Incidentally, casting a particular game into supergame matrix form is generally not a feasible technique of analysis; the number of rows and columns (i.e., the number of sequential-move strategies) becomes astronomically large, even for quite simple games. To illustrate, consider a 3×3 matrix, with Column to choose first; add a prior opportunity for Row to commit himself to any partially or fully specified strategy of response; finally, to study the "defense" against threats, allow Column a still earlier opportunity to commit his choice of column. That is, Column may first commit himself unconditionally if he pleases, Row may then commit himself conditionally in whatever way he pleases, then Column chooses a column and finally Row chooses a row. Let us not complicate the game by limiting sizes of penalties or by inserting any uncertainty or imperfect communication system. This "simple" game, which is not terribly difficult to analyze in its extensive form, turns out to have more than a "googol" (1 followed by a hundred zeros) of columns.

[30] It also explains why a "promise" to abstain from a choice that would damage the other player may not be welcomed by him. A promise that *permits* him safely to make a particular choice may assure us that he *would* make it,

immunizing the owners against extortionate threats against their relatives still in China. Quite likely, for owners located in the United States, the very penalties on transfer of funds to Communist China enhanced their capacity to resist extortion. Deliberately putting one's own assets in a form that made evasion of the law more difficult, or lobbying for more severe penalties on illegal transfer of one's own funds, or even getting one's self temporarily identified as a Communist sympathizer so that his funds would be blocked might have been an indicated tactic for potential victims, to discourage the extortionate threat in advance.

A similar principle is reflected in Article 26 of the Japanese peace treaty, which gives the United States certain claims if subsequent Japanese territorial concessions to other powers are more favorable. When the Japanese were reported to be under pressure from the Russians for additional territorial concessions in 1956, Secretary of State John Foster Dulles (9) pointedly described that article of the treaty in his press conference and said that he had recently "reminded the Japanese of the existence of that clause." The evident intention was to strengthen Japanese resistance; and it may be supposed that by "reminding" the Russians of the same clause through the medium of his press conference, Dulles helped to provide the Japanese with the familiar bargaining claim, "If I did it for you, I'd have to do it for everyone else." It was, in terms used earlier, a "commitment" secured by the penalty of a forfeit to the United States. (Paradoxically, the United States could not give the Japanese the benefit of this bargaining gimmick unless the United States were patently motivated to take advantage of its claim if the tactic failed.)

so that we can count on it and make some prior choice that is to his disadvantage. By the same token, *adding* values selectively to the other's payoffs can absolutely worsen his position—if we have a means of making the addition. In the accompanying matrix, assuming Row has first move, Row can "win"—he can gain 7 at Column's expense—if he unilaterally guarantees to compensate Column in the event of an outcome at i,II, the compensation coming out of his own winnings. If he promises to pay 2 to Column in such an event, he gets 8; Column gets 3; otherwise, without the promised compensation, Row cannot choose i, and the outcome is at ii,I with payoffs of 1 and 10, respectively. Column obviously prefers that Row be unable to commit himself to confer the "benefit." (If the blackmailer cannot scale down his demands to where what he demands, plus the fine for paying blackmail, are less than the damage he threatens, he may offer to pay his victim's fine. This guarantees what his victim's response to the threat will be; so the threat is made, to the disadvantage of the victim.)

0	2	10	1	becomes	0	2	8	3
1	10	2	0		1	10	2	0

"STRATEGIC MOVES"

If the essence of a game of strategy is the dependence of each person's proper choice of action on what he expects the other to do, it may be useful to define a "strategic move" as follows: A strategic move is one that influences the other person's choice, in a manner favorable to one's self, by affecting the other person's expectations on how one's self will behave. One constrains the partner's choice by constraining one's own behavior. The object is to set up for one's self and communicate persuasively to the other player a mode of behavior (including conditional responses to the other's behavior) that leaves the other a simple maximization problem whose solution for him is the optimum for one's self, and to destroy the other's ability to do the same.

There is probably no contrast more strik-

ing, in the comparison of the mixed-motive and the pure-conflict (zero-sum) game, than the significance of having one's own strategy found out and appreciated by the opponent. Hardly anything captures the spirit of the zero-sum game quite so much as the importance of "not being found out" and of employing a mode of decision that is proof against deductive anticipation by the other player.[31] Hardly anything epitomizes strategic behavior in the mixed-motive game so much as the advantage of being able to adopt a mode of behavior that the other party will take for granted.

It can, of course, be an advantage in the zero-sum game to have the opponent believe firmly in a particular mode of play for one's self, but only if that belief is in error. In the mixed-motive game, one is interested in conveying the *truth* about his own behavior—if, indeed, he has succeeded in constraining his own behavior along lines that, when anticipated, win.

Another paradox of mixed-motive games is that genuine ignorance can be an advantage to a player if it is recognized and taken into account by an opponent. This paradox, which can arise either in the coordination problem or in the immunity from a threat, has no counterpart in zero-sum games.

A related contrast between zero-sum and mixed-motive games is that the former, characterized by secrecy and purely unilateral decision processes, can, with reason, be viewed as having rational solutions that are necessarily symmetrical, while the mixed-motive game, if it has strategic moves, is inherently asymmetrical. The richness of the game in its strategic potentialities depends on the fact that all moves are not foredoomed to neutralization by a symmetrical move structure stamped on a game by definition and known to the players as inevitable. (This point concerning symmetry is argued at length in an earlier paper by the author [32].)

III. *Co-ordination of Expectations in the "Pure"-Bargaining Game*

There remains the tantalizing case of the "pure"-bargaining game—the game in which, to express it in abstract form, the two or more participants must overtly agree on the division of a sum of gains, or forego the gains altogether—with no obstacles to communication, no secrets, and no "moves" but sheer negotiation. The gains available for distribution may or may not be completely divisible; they may or may not be of similar utility to the two parties; but, among the array of potential efficient outcomes, more for one means less for the other, and they both know it. This kind of game is of special analytical importance because some element of it is present in any bargaining game that does not resolve itself in a sequence of purely tactical moves, i.e., moves (of the kind discussed in Part II of this paper) that make the outcome formally determinate; and this element of "pure" bargaining is in fact the most tantalizing part of non-zero-sum game theory.[32]

---

[31] Concerning this point, Von Neumann and Morgenstern say (38, p. 147): "We have placed considerations concerning the danger of having one's strategy found out by the opponent into an absolutely central position."

[32] It should be emphasized that a solution to the "pure"-bargaining problem cannot necessarily be superimposed on, say, the Von Neumann–Morgenstern "solution," as though it applied to the residual "range of indeterminacy" left by another theory. Consider the real estate example mentioned in footnote 16 on p. 222 and suppose that B does not yet exist. *Something* determines a price between 10 and 25. Suppose that the "something," in a particular case, would have led to a price of 12. B appears, approaches C,

In a two-person model, this game can be represented by a set of points in an *X*, *Y* plane, one of which is to be chosen by agreement between players X and Y, to yield rewards denoted by the *X*, *Y* co-ordinates, plus a zero point that obtains by default in the absence of an agreed choice. We now have a game in which any outcome is a point from which at least one party would be willing to retreat for the sake of agreement, and the other knows it. For a party to insist on any such point (other than his least favorite among the efficient points) is to take a pure "bargaining position," since one always *would* take less rather than reach no agreement at all and since one always *could* recede if retreat proved necessary to agreement. The only reason for accepting any point short of one's most favored point is the expectation that the other will not concede more; yet there is no logical reason for anybody to expect anything except what he expects to be expected to expect. Agreement is reached when the two parties' expectations converge on a particular point in such fashion that neither expects the other to concede anything further, knowing that whether or not the other will concede depends simply on what the other expects to be expected to expect. Somehow the infinitely reflexive expectations have to converge on a particular outcome that both parties have come to consider inevitable.

What causes their expectations to converge on some resting place in this fashion? What causes some particular outcome to acquire this aura of finality? What makes each recognize that they both recognize that bargaining has reached an end? And how do we discover what characteristics of a particular outcome make it eligible for this sort of election? How do we discover what kinds of tactics can lead to this point of common consent?

The author (31) argued in an earlier paper that what brings expectations into convergence and brings the negotiations to a close is the "intrinsic magnetism" of particular outcomes, especially those that enjoy prominence, uniqueness, simplicity, precedent, or some rationale that makes them qualitatively differentiable from the continuum of possible alternatives. But what is "intrinsic magnetism"? In rationalizing this idea (which seems supported by empirical evidence, some of which was mentioned in that article) it was proposed that there is often a close analogy between the outcome of an *explicit* bargain and the outcome that would seem most plausible if agreement had had to be reached by a *tacit* process. In tacit bargaining the parties have *to let themselves be co-ordinated* toward a common choice, much as they would follow the lead of an arbitrator in search of

---

and threatens to offer A 13. What is it worth to C to prevent this?

Something must have convinced A that he could not get more than 12 (even, under the strict terms of the theory, knowing that C's reservation price was 25). B can destroy that conviction; he can disturb expectations; he can force C to contradict his earlier behavior, change his role, break his pattern of bargaining, and begin anew after the shock. Who knows what price will be settled on next time—what price the "something" of our unspecified theory will determine with B in there just messing around. The damage to C that B can cause, just by participating in the bargaining process, gives him the basis for an extortionate threat. And there is no a priori reason to suppose that C's judgment of what it is worth to keep B from stirring things up is limited by how badly B might want the house, especially when B's interest is concentrated on getting money, not the house, anyway. Thus the limits set by the original "solution," which left a residue of indeterminacy, may be compromised by the theory that one has to cultivate for dealing with that residue.

the agreement that they both want. The conflict of interest was dominated by the need for co-ordination.

The argument was then advanced that a similar psychic phenomenon may occur even when communication is present. Explicit bargaining, like tacit bargaining, requires some "co-ordination" of the participants' expectation of the outcome. It was argued that if in a given game context the participants are intellectually capable of recognizing some unique outcome that can command their agreement through its power to co-ordinate their choices, they will probably be conscious of the focal properties of such an outcome even when they are not dependent on it for their co-ordination. That is, the kinds of properties that would make a particular outcome a strong candidate for election in case of tacit bargaining are likely to be the kinds of properties that have a strong appeal for attention when an explicit agreement is being sought.

Furthermore, the recognition of such a particular outcome by one party implies his recognition that the other can recognize it, too, and his recognition that the other recognizes that the first recognizes it, since in the tacit case the only virtue of such a solution is that each side can recognize that they both will recognize it as inevitable. Neither can disguise or suppress his anticipation because the whole basis for it is the recognition of it as a shared expectation. One's consciousness of the focal properties of such an outcome in the case of *explicit* bargaining is a consciousness that both are conscious of it. There is, so to speak, involuntary tacit consent that both recognize the attraction of this particular point. What was vaguely referred to as "intrinsic magnetism" is therefore not a mystical concept but can be rationalized as an intellectual phenomenon experimentally identifiable in games that demand co-ordination of strategies. The co-ordination of expectations in explicit bargaining is at least *analogous* to the co-ordination of strategies when communication is cut; and it may be the *same* intellectual or psychic phenomenon.[33]

But, although the analogy with tacit bargaining may exercise this influence on the participants, it is by no means unique in doing so; in many cases the tacit-bargaining analogue would not itself have a very powerful solution. This is consequently *an* influence, not the sole influence, on the mutual choice in a pure-bargaining game. What other influences are there that can focus expectations ultimately on some particular outcome and that are analyzable within the framework of game theory rather than dependent on the whole field of psychology?

The writer suggests a generalization of the phenomenon represented in the tacit-bargaining analogy. It is that in these completely "indeterminate" situations, in which the logical structure of the game provides no rationale for the convergence of expectations anywhere, any mutually perceived analogue may exercise a power of suggestion that can focus expectations and thereby bring the players under its discipline. And this analogue may simply be an al-

[33] It should be added that the "intrinsic magnetism" of particular outcomes in a bargaining situation or in a pure co-ordination problem gets some support and clarification from the very substantial body of experimental evidence provided by the Gestalt psychologists. Their work on the perception of physical forms is especially pertinent. For example, incomplete shapes were shown to people whose vision was damaged in part of the eye, and they often saw the shapes as complete rather than as partial. But the particular shapes that they "completed" for themselves followed certain principles of simplicity; and unfamiliar "simple" figures were completed

ternative game—that is, an alternative bargaining situation—that differs in having some prescribed set of moves or communication channels that would make it a determinate game. In other words, if there is a *variant* of the game in question that would have a determinate solution, that solution of the variant game provides a focal point for the convergence of expectations in the indeterminate game in question.

---

where very familiar, but less simple, figures were not. Koffka (19, pp. 140–42, 146–47, 109–10) refers to "spontaneous organization in simple shapes." We are surrounded by skewed rectangles; but what we "see" about us is rectangles, not departures from perfect rectangles, because "the true rectangle is a better organized figure than the slightly inaccurate one would be." Adverting to the minimum-maximum properties of stationary processes, Koffka suggests that psychological processes will have these properties: "For we can at least select psychological organizations which occur under simple conditions and can then predict that they must possess regularity, symmetry, simplicity. This conclusion is based on the principle of isomorphism, according to which characteristics of the physiological processes are also characteristic aspects of the corresponding conscious processes." And, "Thus we have gained a general, though admittedly somewhat vague, principle to guide us in our investigation of psychophysical organization. . . . The principle . . . can briefly be formulated like this: psychological organization will always be as 'good' as the prevailing conditions allow. In this definition the term 'good' is undefined. It embraces such properties as regularity, symmetry, simplicity and others which we shall meet in the course of our discussion."

If individual perception and "organization" of forms follow these constraints, the process of "mutual perception" and "mutual organization of forms" involved in the convergence of expectations must depend on similar restraints at least as rigorous. And, since the non-zero-sum game requires some ultimate joint "organization of form," so to speak, a normative theory of strategy (not just a descriptive psychology) must take these restraints into account.

AN ILLUSTRATIVE GAME

An example—which here depends entirely on the reader's intuitive appreciation of it but should be susceptible of experimental verification—will make the point clear. Suppose two players, X and Y, are to pick jointly a point within or on the boundary of the segment of a circle *OPQ* shown in Figure 10; if they can agree on such a point, they receive sums of money corresponding to the *X* and *Y* co-ordinates, and if they reach no agreement they get zero apiece. Here is the pure-bargaining game. Now let us give specific form to the way in which

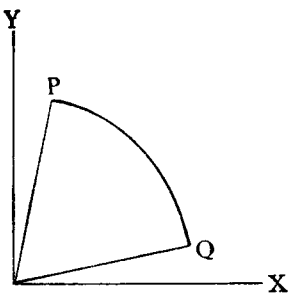

FIG. 10

they will *express* these choices. X and Y are to choose, respectively, the two components of a vector from the origin, namely, the direction and the length. X, so to speak, will point the steering wheel, and Y will work the accelerator; or X will aim the gun horizontally, and Y will control its elevation. Mathematically speaking, we have introduced nothing of substance into the game, since *explicit agreement is required* or they both get nothing. The logical structure of the game is unaffected by how we require the two participants to translate the point they choose into a co-ordinate system or otherwise to identify it for us. But the details we introduce may have connotative significance and may influence the players'

expectations of the outcome and hence the outcome itself.

Note that the particular topical content that we have introduced into the game is asymmetrical; the roles of the two players—or rather the roles that seem to be implied for the two players—and the manner in which they must divide between them the communication of their choice to us differ in a way that *could* have logical significance. X makes a distributive choice; in giving the direction of the vector, he indicates in what proportions he and his partner will divide whatever sum they ultimately get. Y is left responsibility for the efficiency of the solution; he, in effect, appears to decide how near to the boundary the point will be, that is, how large a joint sum they will be dividing between them.

The reason we say that this difference in their *apparent* roles *could* have logical significance is that there are very similar games in which they *would* have logical significance and would make the game quite determinate. Let us look at some alternative games. One variant would have them choose separately without communicating; in this tacit game Y seems to be in a completely subordinate position. He is responsible for a choice that affects him and his partner together; Y alone is responsible for how the gain is divided. Whatever X does, Y can only benefit by choosing the boundary curve; X is free to secure the proportion denoted by Q, hence the amount denoted by Q.

Suppose, instead, they take turns moving, without prior agreement. If X can move first, he points the vector toward Q; again Y has lost and can only lay the point down on the boundary at Q. Now suppose Y gets to move first: he loses again; whatever length he chooses, X will distribute it in the ratio denoted by Q, and Y can gain only by choosing the boundary curve. If, in still another variant, X can commit himself to a choice, he again wins as though he had first choice. But if, alternatively, Y is able to commit himself, it does him no good; the commitment is only a means of acquiring first choice, which, as we saw, does him no good. Y could win only if he (not X) could commit himself to a conditional choice—out on the boundary if X chooses favorably for Y, or all the way back at zero if X chooses unfavorably—i.e., if he could threaten mutual destruction in order to win. The threat is the only tactic in which even an asymmetry in Y's favor can do him any good. (Even the incentive to cheat on an agreement would always be in X's favor, never in Y's.)

Now let us go back to the original game, in which there is none of this superstructure about order of moves, etc. Suppose X declares a direction for the vector: he aims it at Q. This is his offer, and Y says, "No." X says, "That's it," and writes it out on the sheet of paper on which he is to transmit his part of the joint choice. He simply says that he has no intention of changing it, that there is nothing Y can do but accept it, and that if Y chooses anything but the boundary point at Q he will only hurt himself as well as X. X sits in a chair on the other side of the room and says he does not plan to get up until Y has written his own part of the bargain and the game is over.

What can Y do? First, note that Y has no comparable tactic. Y cannot write down a number that favors himself and complacently cross the room and light a cigar. He cannot even pretend that he has made his "final" choice, unless he has already written 100 per cent as his choice of length. Thus in some mechanical sense X has a more "stable" bargaining position than Y. X can create a status quo by declaring what he proposes to choose. Declaring does not constitute a commitment—our rules enforce no

commitment. Y can still threaten non-agreement, to force X to get up and change the direction of his choice more toward *P* or, for that matter, all the way to *P*. But the *roles* of the two players are different. X has a complacent role; he can announce and sit still. Y has to make threats and demands. If X could leave the room, he could tell Y he did not plan to come back, having made his choice. Y could not leave the room and say that he was not coming back, until he had made the choice that satisfied X. (This tactic would be a means of commitment, and so is against the rules.)

Now it should be emphasized that there is nothing in the logic of the situation that gives X any advantage over Y. It remains true that, until the two of them reach explicit agreement, they both stand to get nothing; and Y's ability to withhold agreement until X offers a satisfactory outcome is not in any logical way diminished by comparison with what it would be if their mode of expressing their choice had never been prescribed. It just looks as though X had what would be described as a strong psychological advantage. This advantage bears no relation to the mathematical structure of the game and has no meaning within any mathematical formulation of the problem. Nor is it a "tactical" advantage, of the kind involved when one party can make a threat or a commitment or take advantage of discriminatory communication channels. It is, in a strict sense, psychological. It is created by some topical content that is mathematically inessential to the game. It arises in the power of suggestion that resides in certain procedures, details, or connotative superstructure. Essentially, it resides in the fact that X and Y both perceive that in some way X has an undeniable psychological advantage.[34]

The fact that the advantage rests on intuition is not an objection. We are dealing with a process that involves intuition; and the argument here is that intuition in game situations may rest on something that is objectively analyzable. It may be the power

[34] In particular, it should be noted that the outcome of the game proposed here would violate any axiom of symmetry. The abstract version of the game is completely symmetrical; the two players have identical capacities to withhold action until explicit agreement is reached. The mechanical asymmetry would be a logical part of the abstract game only if there were limits on communication or a prescribed order of moves or if players could commit themselves to their choices before reaching joint agreement. In another paper the author (32) has argued that the "pure" (moveless) explicit bargaining game of the type analyzed in this section may actually not "exist." The point of the argument is that a well-defined game must have a more explicit legal structure than the game informally characterized above on p. 249; in particular, there must be an explicit formula or rule for termination, and there must be an operational definition of what constitutes "agreement" during the game or at its closure. The problem of making up such a game is that, when attention is given to the legal details, explicit-communication games often appear to give way, by default, to a tacit version at some "last moment" unless agreement already exists; and if, then, that tacit version has an efficient solution, it becomes a minimax strategy for a player to force the game into its tacit stage rather than reach prior agreement. If the tacit game has no compelling solution, explicit bargaining acquires the tactical significance of moves designed to create focal points for co-ordination in the final stage, each player interested in creating focal points favorable to himself. The reader can experiment with these possibilities by considering precisely how the game discussed in the text might be given rules sufficiently explicit to permit actual play, i.e., rules for termination (by time limit, roll of dice, etc.) and for establishing whether a legal contract exists at termination or sooner. From this point of view it is not quite fair to say that the nominal asymmetry in the game discussed in the text is not part of the logical structure: we have not given the game a closed logical structure yet. (We have left it in the form in which game theory traditionally leaves it.)

of suggestion of analogous games: the ability of the participants to recognize—or, rather, their inability not to recognize—that various details of their situation and the roles they play point toward particular outcomes, because in situations only slightly different they would *determine* those outcomes. Intuition that is governed by analogy, when the analogues are objectively analyzable, is analyzable.

Of course, we must avoid assuming that everything the analyst can perceive is perceived by the participants in a game. The proposition advanced here relates only to the appreciation by actual participants that certain outcomes would be "obvious" in powerfully similar situations—conceivably, even situations that impress their similarity so strongly on the players that the latter do not consciously recognize the difference.

For this reason, game analogues that have sophisticated mathematical solutions (except when the same solution can also be reached by an alternative, less sophisticated route) would not have this power of focusing expectations and influencing the outcome. Or, rather, they would have this power only if the players perceived each other to be mathematicians. Here is the empirical interpretation of such "solutions" as those of Braithwaite, Nash, Harsanyi, and others. It is that the mathematical properties of a game, like the aesthetic properties, the historical properties, the legal and moral properties, the cultural properties, and all the connotative details, can serve to focus the expectations of certain participants on certain solutions. If X and Y in the above game were mathematical game theorists, they might mutually perceive and be powerfully affected by potential solutions that had compelling mathematical properties.[35] Each would transcend, and know that the other would transcend, such adventitious details as how they were to express their choice. Their analogue would not necessarily be the same one that influences non-mathematician players of the same game.

Thus mathematical solutions are one species of a genus of influences that have the power to focus expectations; but they work through the same psychic mechanism—this power of suggestion that is able to bring expectations into convergence—as the other species. When husband and wife, separated in a department store, gaily traipse off to the Lost and Found by a tacit and jocular mutual appreciation that it is the "obvious" place to meet, two mathematicians in the same situation—each aware that both are aware that both are mathematicians—might look for a geometrically unique point rather than one that depended on a play on words.[36]

The nature of convergent expecta-

---

[35] In many cases, though, these properties would be a uniqueness or symmetry that would have non-mathematical definitions and non-mathematical appeal, too, or would happen to coincide with qualitatively distinguishable points that could be rationalized in an equally compelling non-mathematical way.

[36] The main point here is independent of whether, under the "rules" of game theory, a rational player must be presumed to know as much mathematics as he ever has need for. We are dealing here with the players' shared appreciations, preoccupations, obsessions, and sensitivities to suggestion, not with the resources that they can draw on when necessary. If the phenomenon of "rational agreement" is fundamentally psychic—convergence of expectations—there is no presumption that mathematical game theory is essential to the process of reaching agreement, hence no basis for presuming that mathematics is a main source of inspiration in the convergence process. The writer (32) has tried to argue that no *efficient* outcome of a bargaining game, however it divides the gains between the players, can constitute refutation of the rationality hypothesis as long as rationality is not given a definition implicit in a particular theory.

tions. One may or may not agree with any particular hypothesis as to how a bargainer's expectations are formed, either in the bargaining process or before it and either by the bargaining itself or by other forces. But it does seem clear that the outcome of a bargaining process is to be described most immediately, most straightforwardly, and most empirically in terms of some phenomenon of stabilized convergent expectations. Whether one agrees explicitly to a bargain or agrees tacitly or accepts it by default, he must, if he has his wits about him, expect that he could do no better and recognize that the other party must reciprocate the feeling. Thus the *fact* of an outcome, which is simply a co-ordinated choice, should be analytically characterized by the notion of converging expectations.

The intuitive formulation, or even a careful formulation in psychological terms, of what it is that a rational player "expects" in relation to another rational player in the "pure" bargaining game, poses a problem in sheer scientific description. Both players, being rational, must recognize that the only kind of "rational" expectation that they can have is a fully shared expectation of an *outcome.* It is not quite accurate—as a description of the psychological phenomenon—to say that one expects the second to concede something or to accept something; the second's readiness to concede or to accept is only an expression of what he expects the first to accept or to concede, which, in turn, is what he expects the first to expect the second to expect the first to expect, and so on. To avoid an "ad infinitum" in the descriptive process, we have to say that both sense a shared expectation of an *outcome;* one's expectation is a belief that both identify the *same* outcome as being indicated by the situation, hence as virtually inevitable. Both players, in effect, accept a common authority—the power of the game to dictate its own solution through their intellectual capacity to perceive it—and what they "expect" is that they both perceive the same solution.

Viewed in this way, the intellectual process of arriving at "rational expectations" in the full-communication "pure"-bargaining game is virtually identical with the intellectual process of arriving at a co-ordinated choice in the tacit game. The actual solutions might be different because the game contexts might be different, with different suggestive details; but the intellectual nature of the two solutions seems virtually identical, since both depend on an agreement that is reached by tacit consent. This is true because the explicit agreement that is reached in the full-communication game corresponds to a priori expectations that were reached (or in theory could have been reached) jointly but independently by the two players before the bargaining started. And it is a tacit "agreement" in the sense that both can hold confident rational expectations only if both are aware that both accept the indicated solution in advance as *the* outcome that they both know they both expect.

A hypothetical experiment. As an illustration of what the author has in mind, the following hypothetical experiment can be considered. (Hopefully, some such experiment could be carried out.) It is offered here as a conceptual analogue or, conceivably, an empirical test of the psychic phenomenon involved in bargaining.

The first stage in the experiment is to invent a machine, perhaps on the principle of the lie detector, that will record or measure a person's "recognition" or the focus of his attention or his alertness or his excitement. What we want is a machine that measures, as the player scans an array of possible outcomes in some orderly fashion, the extent to which particular outcomes

catch his attention or generate excitement in the course of actual bargaining.

Given the machine, set up a bargaining game. For simplicity, make it one in which there are certain gains to be shared when agreement is reached on the shares. Give the game enough "topical content" to provide some room for argument, casuistry, alternative rationales, etc.; that is, provide more than a bare mathematical range with a conspicuous mid-point.

Now have the two players connected to their machines in such a way that each can see the meter on his own machine, each can see the meter on the other's machine, and each is aware that both are aware that both can see both meters. In other words, they mutually perceive that they both can see each other's reactions to particular outcomes as they come within view of the scanning device. We employ a mechanical scanning device, which moves about in the range of possible outcomes, pointing to, lighting up, or focusing on one possible outcome after another. It follows perhaps some regular course, perhaps a random course. Let this machine scan; let the players watch it scan, watch their own and each other's meters, and watch each other's faces if they wish to.

Finally, we go through with the game; and there may be several variants. An interesting possibility would be to exclude explicit bargaining and simply let the scanning proceed, back and forth or round and round among the array of alternative outcomes. We watch to see whether the recorded reactions of the two players tend eventually to converge on a single outcome, in the sense that their involuntary, physically identifiable reactions are at some kind of maximum for the same particular outcome among all those to which the scanning device elicits their reactions. (For control purposes, we might once have subjected each player to a scanning session in which the other player was absent, to get some notion of each player's reactions independently of any interaction between the players.) If convergence does occur, we have certainly identified a significant phenomenon, whether or not we can allege that this is *the* psychic bargaining process. We shall have demonstrated (*a*) that players do react to the content of the bargaining situation and (*b*) that their reactions are subject to a mutual interaction that results from the fact that each can see the other's reaction and each knows that his own visible reaction is yielding information about his own expectations. (The writer conjectures that, like Lot's wife, players will often be unable to keep their attention from being drawn to particular outcomes, even unfavorable outcomes, and that a conscious effort to ignore a "focal point" may often enhance the focal power.)[37]

Another variant would be to let the players bargain explicitly during the scanning and metering, with the scanning device inexorably eliciting their physical reactions in the course of the discussion in a

[37] The following observation, quoted by K. Koffka (19), may be hard to believe but is certainly to the point: "When an expert . . . follows a football game attentively he will also notice that the goalkeeper, standing before the comparatively large goal, is more often hit than can be accounted for by the mere adventitious kicking of the contestants, even when one takes account of the fact that the goalkeeper whenever he can will try to intercept the ball. The goalkeeper furnishes a prominent point in space which attracts the eyes of the opposing kickers. If the motor activity takes place while the kicker's eye is fixed on the goalkeeper, then the ball will generally land near him. But when the kicker learns to reconstruct his field, to change the phenomenal 'centre of gravity' from the goalkeeper to another point in space, the new centre of gravity will have the same attraction as the goalkeeper had before."

manner visible to both of them. (We could even, in this latter case, let a player adduce the evidence of the visible reaction meters if he wished to as a bargaining tactic, pointing out to his partner, for example, that the latter "obviously" cannot expect to hold out for, say, the $60 he is verbally demanding when it is clear from his blood pressure that his mind is settled on $40.)

This experiment would rest on three hypotheses. First, that an individual player would have physically identifiable "reactions" upon contemplating different alternatives among the range of possible game outcomes and that these reactions would be conspicuously different among the different alternatives. Second, that these reactions, when the player knows that they are naked to his partner's eye, would behave in a manner suggestive of bargaining; i.e., that the reactions of the two players, when visible to both of them, would interact in a kind of "bargaining process." Third, that this measured phenomenon, which we liken to a bargaining process, is part of, or is involved in, or is related to, *the* bargaining process as defined in the ordinary way. (An experiment of the sort described might prove especially interesting for the case of more than two persons.)

The experiment has not been carried out and is not adduced as evidence. It has been described here in order to give an operational representation of the theoretical system that the author has in mind in referring to the "convergence" of expectations and to suggest that the convergence that ultimately occurs in a bargaining process may depend on the dynamics of the process itself and not solely on the a priori data of the game.

Some dynamic characteristics of focal-point solutions. The dependence of a "focal-point" solution on some characteristic that distinguishes it qualitatively from the surrounding alternatives has important dynamic considerations. For example, it often makes small concessions less likely than large ones; it often means that the focal point is more persuasive as an *exact* expected outcome than as an approximation. If a bargainer has persistently been unsuccessfully demanding 50 per cent, compromise at 47 per cent is unlikely; the small concession may be a sign of collapse. Qualitative principles are hard to compromise, and focal points generally depend on qualitative principles. One cannot expect to satisfy an aggressor by letting him have a few square miles on this side of a boundary; he knows that we both know that we both expect our side to retreat until we find some persuasive new boundary that can be rationalized.

In fact, a focal point for agreement often owes its focal character to the fact that small concessions would be impossible, that small encroachments would lead to more and larger ones. One draws a line at some conspicuous boundary or rests his case on some conspicuous principle that is supported mainly by the rhetorical question, "If not here, where?" The more it is clear that concession is collapse, the more convincing the focal point is. The same point is illustrated in the game that we play against ourselves when we try to give up cigarettes or liquor. "Just one little drink," is a notoriously unstable compromise offer; and more people give up cigarettes altogether than manage to reach a stable compromise at a small daily quota. Once the virgin principle is gone, there is no confidence in any resting point, and expectations converge on complete collapse. The very recognition of this keeps attention focused on the point of complete abstinence.

Sometimes the focal point itself is inherently unstable. In that case it serves not as an outcome but as a sign of where to

look for the outcome. This is often true of a "test vote" in a legislative body or a "test issue" that arises in the relations between the players in some continuing game. Often it is a challenge or a dare or an act of defiance that, by its nature, must either elicit a submissive response from the other party or be submissively withdrawn. It is a small piece of the game that comes to symbolize the game itself, setting a pattern of expectations that extends beyond the substance of the point involved. Sometimes it is so intended and constitutes a deliberate tactic; in other cases the act or the issue develops an unintended symbolic significance, making compromise impossible.

Diplomatic recognition of the Communist regime in China, loyalty oaths at universities, a strike settlement in a key industry, surrender of the floor to an interrupter at a cocktail party, or the vote on some particular motion at a political convention may all have this kind of significance. Sometimes, it is true, the outcome on this particular issue simply yields evidence of how other issues would be decided, as when a test vote indicates exactly how large the opposition to a measure is; but often the particular issue is not representative of the rest of the game, it just acquires tacit recognition as a clue to all that will follow, so that each side is the prisoner or beneficiary of the mutual expectations that are created.

Often this phenomenon can be identified as an actual signal in a co-ordination game. The members of an unorganized coalition can often recognize the potentialities of concerted action without being sure that "agreement" exists to act in concert. One wants to know how everyone else is going to act and whether everyone else will do what he knows he ought to. A test vote in a legislature or some particular simultaneous action among the group, like a mass protest, is often a means of "ratifying" the existence of the coalition and of demonstrating that everybody expects everybody else to act in concert. But even in a two-person game, as typified by the dare, the phenomenon of psychological dominance or submissiveness may prove to be psychologically identical with the resolution of a bargaining game.

This process, by which particular moves in a game or offers and concessions achieve symbolic importance as indicators where expectations should converge in the rest of the game, seems to be an area in which experimental psychology can contribute to game theory.

IV. *Game Theory and Experimental Research*

The foregoing discussion suggests several conclusions about the methodology appropriate to a study of bargaining games. One is that the mathematical structure of the payoff function should not be permitted to dominate the analysis. A second one, somewhat more general, is that there is a danger in too much abstractness: we change the character of the game when we drastically alter the amount of contextual detail that it contains or when we eliminate such complicating factors as the players' uncertainties about each other's value systems. It is often contextual detail that can guide the players to the discovery of a stable or, at least, mutually non-destructive outcome. In terms of an earlier example, the ability of Holmes and Moriarty to get off at the same station may depend on the presence of something in the problem other than its formal structure. It may be something on the train or something in the station, something in their common background, or something that they hear over the loudspeaker when the train stops; and, though it may be difficult to derive scientific generalizations about what it is that serves their need for co-

ordination, we have to recognize that the *kinds* of thing that determine the outcome are what a highly abstract analysis may treat as irrelevant detail.

A third conclusion, which is particularly applicable whenever the facilities for communication are short of perfect, where there is inherent uncertainty about each other's value systems or choices of strategies, and especially when an outcome must be reached by a sequence of moves or maneuvers, is that some *essential* part of the study of mixed-motive games is necessarily empirical. This is not to say just that it is an empirical question how people do actually perform in mixed-motive games, especially games too complicated for intellectual mastery. It is a stronger statement: that the principles relevant to *successful* play, the *strategic* principles, the propositions of a *normative* theory, cannot be derived by purely analytical means from a priori considerations.

In a zero-sum game the analyst is really dealing with only a single center of consciousness, a single source of decision. True, there are two players, each with his own consciousness; but minimax strategy converts the situation into one involving two essentially unilateral decisions. No spark of recognition needs to jump between the two players; no meeting of minds is required; no hints have to be conveyed; no impressions, images, or understandings have to be compared. No social perception is involved. But in the mixed-motive game, two or more centers of consciousness are dependent on each other in an essential way. Something has to be communicated; at least some spark of recognition must pass between the players. There is generally a necessity for some social activity, however rudimentary or tacit it may be; and both players are dependent to some degree on the success of their social perception and interaction. Even two completely isolated individuals, who play with each other in absolute silence and without even knowing each other's identity, must tacitly reach some meeting of minds.

There is, consequently, no way that an analyst can reproduce the whole decision process either introspectively or by an axiomatic method. There is no way to build a model for the interaction of two or more decision units, with the behavior and expectations of those decision units being derived by purely formal deduction. An analyst can deduce the decisions of a single rational mind if he knows the criteria that govern the decisions; but he cannot infer by purely formal analysis what can pass between two centers of consciousness. It takes at least two people to test that. (Two analysts can do it, but only by using themselves as subjects in an experiment.) *Taking a hint* is fundamentally different from deciphering a formal communication or solving a mathematical problem; it involves discovering a message that has been planted within a context by someone who thinks he shares with the recipient certain impressions or associations. One cannot, without empirical evidence, deduce what understandings can be perceived in a non-zero-sum game of maneuver any more than one can prove, by purely formal deduction, that a particular joke is bound to be funny.

To illustrate, consider the question whether two people, looking at the same ink blot, can identify the same picture or suggestion in it if each is trying and knows that the other is trying to concert on the same picture or suggestion? The answer to this question can be found only by trying. But if they can, they can do something that no *purely formal* game theory can take into account; they can do *better* than a purely deductive game theory would predict. And if they can do better—if they can rise above

the limitations of a purely formal game theory—even a normative, prescriptive, strategic theory cannot be based on purely formal analysis. We cannot build either a descriptive theory or a prescriptive theory on the assumption that there are certain intellectual processes that rational players are *not* capable of, of the kind involved in "taking a hint"; it is an empirical question whether rational players, either jointly or individually, can actually do better than a purely formal game theory predicts and should consequently ignore the strategic principles produced by such a theory.[38]

Again it should be emphasized that the reason why this kind of consideration does not arise in the zero-sum game is that any such social interaction could not be to the advantage of both players simultaneously and that at least one of the rational players would have both motive and ability to destroy all social communications. But in a non-zero-sum game that involves any initial uncertainty over which among the possible outcomes are in fact efficient and any need for co-ordinated mutual accommodation to get to an efficient outcome, a rational player cannot absent himself in self-defense from the social process; he cannot turn off his hearing aid to avoid being constrained by what he hears, if complete radio silence makes efficient collaboration impossible. Nor can he rationally fail to open a letter, once it is delivered, since the other party will have assumed that he will open it and have acted accordingly.

At this point a question arises whether the game-theory trail ramifies indefinitely over the whole domain of social psychology or leads into a more limited area particularly congenial to game theory. Are there some general propositions about co-operative behavior in mixed-motive games that can be discovered by experiment or observation and that yield a widely applicable insight into the universe of bargaining situations? Although success is not assured, there are certainly some promising areas for research; and, even if we cannot discover general propositions, we may at least disprove em-

---

[38] A good laboratory example of the communication-perception part of game strategy is the experiment reported by M. M. Flood (10), who presented his players with a 2×2 non-zero-sum matrix for 100 consecutive tacit plays. The special property of the matrix is that the players can win only by co-operating on a particular cell on each play, but to distribute the winnings for the 100-play sequence they must co-operate on some pattern of alternation among two or more cells that discriminate differently between the two players. And the only means of negotiating over the distribution to be sought and concerting on a pattern of alternating play that achieves it is through the choices they actually make as the play proceeds. This "communication" stage—and any later stage when one player may depart from the tacitly agreed pattern to cheat a little and have to be punished by a reprisal pattern—is jointly expensive to them, since an unco-ordinated choice is a lost chance to make some money.

The question of how to communicate a proposal effectively and how to interpret the other player's proposal implicit in his pattern of play is evidently dependent on some mutual perception of a shared sense of pattern—a jointly recognized ability to complete a pattern of which a fragment has been displayed—not unlike the process involved in the experiments of the Gestalt psychologists mentioned in an earlier footnote. And, while a purely formal theory of communication may derive certain minimum standards of "efficiency" in communication that rational players ought to achieve, it is an empirical question whether players can do better than that. How well one can take a hint and what kinds of hints are most successful are empirical questions of social perception, probably amenable to experimental study. (The same problem arises if two men at an auction recognize that they are jointly losing money by bidding against each other and try, without giving any overt evidence of collusion, to concert on some pattern of reciprocal and alternating abstention from bidding that both saves them money jointly and distributes the savings and the opportunities between them.)

pirically some that are widely held. It does appear that game theory is badly underdeveloped from the experimental side.

Consider a game like the one described earlier, involving the movement of counters over a map, or the modified chess game that was made non-zero-sum. These can be taken to represent games in "limited war"; both players can gain by successfully avoiding mutually destructive strategies. Here is a game in which the ability of the two players to avoid mutual destruction may well depend on what *means* for successful coordination of intentions are provided by the incidental details of the game, by such things as a configuration of the map or board, the suggested names of the pieces, the tradition or precedent that goes with the game, and the scenario or connotative background that is instilled into the players before the game begins. It is a sufficiently complicated game to require perceptive play by both sides and the successful conveyance of intentions. If we suppose for a moment that the technical problem of constructing a playable game of that type has been mastered, it is worth while to consider what line of questions we might try to investigate or what hypotheses we might test.

One such question would be this: by and large, does it appear that the players are any more successful in reaching an efficient solution, i.e., a mutually non-destructive solution, when (*a*) full or nearly full communication is allowed, (*b*) no communication or virtually none is allowed, other than what can be conveyed by the moves themselves, or (*c*) communication is asymmetrical, with one party more able to send messages than he is to receive them? There is no guaranty that a single, universally applicable answer would emerge; nevertheless, some quite general valid propositions about the role of communication might well be discovered. The enormous significance of this question is attested by some of the current controversies about whether the possibility of keeping war limited is greater if there is good communication between both sides, or if there are unilateral declarations ahead of time by one side or the other, or if there is virtually no overt communication between the belligerents.[39]

Another set of questions, also pertinent to problems of limited war, international or other, would be whether a stable, efficient outcome is more likely when the connotations of the game—the names and interpretations that are overtly attached to the moves and pieces and objects on the board—are familiar and recognizable or when they are quite novel, unfamiliar, and unlikely to inspire similar notions in the two players. Is it—to speak of the game in a particular extensive form—more likely that rational players can keep a war limited in Southeast Asia, using conventional and atomic weapons, or in a battle against an unknown adversary on the surface of the moon, using strange bacterial weapons? These are important questions; they are at the very cen-

[39] To preclude any possible misunderstanding: the writer is not suggesting that limited war can be simulated in the laboratory or that experimental results regarding the limiting process can be directly transferred to the outside world. Experiments of the kind described would come under the heading of "basic research." And it would be concerned mainly with the perceptual and communicative side of the problem, not the motivational—except to the extent that motivations affect social perception. The probability that the results of such research would find ready application, however, is enhanced by the observation that much current theorizing on, for instance, the role of communication in limited war or the types of limitations most likely to be observed seems itself to be based only on what might be described as implicit experimental games played introspectively.

ter of game theory; and they are questions that cannot possibly be given a confident answer without empirical evidence. And there is no arguing that rational players have the intellectual capacity to rise above these details of the game and ignore them; the importance of the details is that they can be supremely helpful to both players and that rational players know that they may be dependent on using these details as props in the course of their mutual accommodation.

Is a stable, efficient outcome more likely between two players of similar temperament and cultural background or between two quite different players? Is a stable, efficient solution more likely with two practiced players, two novices, or one novice and a practiced player; and in the latter pair, who has the advantage?

In a game of this sort, how crucial are the opening moves? If stable patterns of behavior, i.e., "rules of the game," are not discovered early, will they be discovered at all? Is mutually successful play more likely if the general philosophy of each player is to begin with "tight" rules or highly "limited" weapons and resources, loosening them a little only as the occasion demands it, or if each player sets himself wider limits at the outset in order to avoid having to establish a practice of loosening rules as he goes?

How much influence on a game of this sort can a "mediator" have, and what kinds of mediating roles are most effective? Does it help or hinder the other two players if the mediator has a stake of his own in the outcome? To what extent can a mediator discriminate in favor of one of the two players and still increase the likelihood of a stable, efficient outcome?

It would be interesting in a game of this sort to have the players score both themselves and their partners from time to time on such matters as who is playing the more aggressively or the more co-operatively, and what "rules" each thinks are in force and thinks the other thinks are in force; of who is "winning" in a bilateral sense (it being recalled that the substantial ignorance of each other's value system makes this always a matter of interpretation); of when the game has reached a "critical" turning point, or when an "innovation" in tactics has been introduced, or when a particular move by the other side is to be interpreted as "retaliation" or a new initiative.

Because a "law of reprisal" is essentially *casuistic* in nature; because the mutually recognized restraints in any form of "limited war" are essentially based on something psychologically and sociologically akin to *tradition;* and because the received body of casuistry and tradition is often wholly inadequate to the game at hand (say, graduated atomic reprisal on the U.S.S.R. and America while limited atomic war obtains in Europe, or the bombing of grammar schools in an area without recent experience in racial violence, or the introduction of new forms of non-price competition in a particular industry), it seems likely that the empirical part of game theory will include experimental work like that of Muzafer Sherif (35) reported in "The Development of Social Norms." He finds that when no norms exist for a laboratory judgment, they are created by the subjects; and when norms are created for two parties in the same process, each player's developing norm influences the other's. There is a process of genuine learning with respect to *values;* each side adapts its own system of values to the other's, in forming its own. When the supply of available "objective" criteria is incapable of yielding a complete set of rules, i.e., when the game is "indeterminate," norms of some sort must be developed, mutually perceived, and accepted; patterns of action and response

have to be legitimized.[40] In an almost unconsciously co-operative way, adversaries must reach a mutually recognized definition of what constitutes an innovation, a challenging or assertive move, or a co-operative gesture, and they must develop some common norm regarding the kind of retaliation that fits the crime when a breach of the rules occurs.[41]

A "scenario" might, for example, identify one of the players as "aggressor"; it might give the outcomes of previous plays of the same game by other players; it might give a background story that would tend to identify some particular division of the terrain as corresponding to an original "status quo"; or it might seem to attach a kind of moral claim of one of the players to particlar parts of the board. These background data would have no influence on the logical or mathematical structure of the game; they would be intended to have no force except power of suggestion. Again, one might set up the board so that on the first play it corresponds to the way it stood in the middle of the same game as played earlier by two other players, and see whether the outcome can be effected by informing the players of what the starting lineup was in that earlier game. If players tend to develop "norms" based on the static configuration of the game as they appreciate it at the outset, it may be possible to distort those norms by providing, in a completely "non-authoritative" way, a background story that suggestively indicates some other hypothetical starting point.[42]

It should also be interesting to see whether each player can really discern when the other is "testing" his determination, "daring" him, and so forth; and it might be possible to study the process by which particular encounters become invested with symbolic importance, such that each player recognizes that he is establishing a role and reputation in the way he conducts himself at a particular point in the game.

Another dimension of the game that seems susceptible of analysis is the significance of the *incrementalism* that is involved in the moves and value systems. Take, for example, a game that involves moving pieces over a board or troops over

---

[40] A splendid example of the creation of norms in practice—and one that suggests that the process is susceptible of analysis—was the acceptance during the 1957 disarmament discussions of the notion that any inspection zone ultimately agreed on had to be selected from among the array of possible pie-shaped zones with apex at the North Pole.

[41] One may hope, as a game theorist, that a clear line can be drawn between the experimental psychology pertinent to game theory and the rest of social psychology; this is still supposed to be a theory of *strategy*, not the entire domain of conflict behavior. But it is not clear just where the line can be drawn in advance. "Hostility," for example, might seem to be an emotional or temperamental quality best kept out of game theory; but if a player's hostility in the game is a significant constraint on his ability to perceive the other player's meaning, it becomes part of the "communication structure." An experiment by Morton Deutsch (6, pp. 113–14) is pertinent. He let pairs of players play non-zero-sum games (in matrix form) tacitly for a sequence of two plays, the game providing both a "co-operative" and an "unco-operative" choice. Those who played unco-operatively against a co-operative partner had an opportunity, on the second play, to respond to the implicit offer of co-operation. But, "when their expectation of the other person's choice was not confirmed, they tended to interpret his choice as being a function of indifference or a basic lack of understanding as to how the game 'should' be played. . . . In this group, knowledge of the other person's choice, because of the meaning attributed to it, tended to reinforce the previous negative sentiments regarding the intentions of the other person."

[42] The income tax games described in the author's earlier paper in this *Journal* indicate the force of this power of suggestion (31, pp. 7–8, 9).

some terrain. If players move in turn, each moving one piece one square at a time, the game proceeds at a slow tempo by small increments; the situation on the board may change character in the course of play, but it does so by a succession of small changes that can be observed, appreciated, and adapted to, with plenty of time for the mistakes of individual players or mutual mistakes that destroy value for both of them to be observed, adapted to, and avoided in subsequent play. If there is communication, there is time for the players to bargain verbally and to avoid moves that involve mutual destruction. But suppose that, instead, the pieces can be moved several at a time in any direction and any distance and that the rules make the outcome of any hostile clash enormously destructive for one or both sides. Now the game is not so incremental; things can happen abruptly. There may be a temptation toward surprise attack. While one can see what the situation is at a particular moment, he cannot project it more than a move or two ahead. There seems to be less chance to develop a modus vivendi, or tradition of trust, or dominant and submissive roles for the two players, because the pace of the game brings things to a head before much experience has been gained or much of an understanding reached. But does a more incremental game make successful collaboration easier, or does it just invite a riskier mode of play? Or does this depend on what kinds of people the players are and on what suggestions we plant in the game itself? Is the critical factor the incrementalism of the *moves* in the game or incrementalism in the *value* systems of the players (i.e., of the scoring system)? Or can these be made commensurate with each other, so that incrementalism can be introduced into a game in one dimension to offset the lack of it in another? The relevance of these questions is attested by the controversy over the role of nuclear weapons in limited war, the significance of the temptation to surprise attack in a situation that depends on mutual deterrence, and various proposals to reduce the tempo of modern war and to isolate it geographically, together with disagreement over whether there can be such a thing as limited war on the continent of western Europe. Incrementalism may be comparatively amenable to formal analysis, once the necessary empirical benchmarks have been identified by experiment or observation.[43]

These questions have concerned two-person games, except for the possible role of the mediator. Similar games could be played by three or more participants, each on his own account; and the author conjectures that—at least among "successful" players—many of the empirical results would appear in sharper relief with the larger number of players. More generally, the kind of co-ordination involved in the formation of mobs and coalitions may lend itself to experimental study. In contrast to the more sanitary, symmetrical schemes that have sometimes been used to study the formation of coalitions in game theory, it might prove more interesting to introduce deliberately certain asymmetries, precedents, orders of moves, imperfect communication structures, and various connotative details, in order to study the crystallization of groups. Certainly the influence exerted on the formation of coalitions by various kinds of asymmetrical and otherwise imperfect

[43] Kissinger (18, p. 225) says: "It is not only that limited war must find means to prevent the most extreme violence; it must also seek to slow down the tempo of modern war lest the rapidity with which operations succeed each other prevent the establishment of a relation between political and military objectives. If this relationship is lost, any war is likely to grow by imperceptible stages into one all-out effort."

communication systems often lends itself to systematic experimental study.[44]

REFERENCES

1. ACHESON, DEAN. *Power and Diplomacy.* Cambridge: Harvard University Press, 1958.
2. BAVELAS, A. "Communication Patterns in Task-oriented Groups." In D. CARTWRIGHT and A. F. ZANDER, *Group Dynamics.* Evanston: Row, Peterson & Co., 1953.
3. BRAITHWAITE, R. B. *Theory of Games as a Tool for the Moral Philosopher.* Cambridge: Cambridge University Press, 1955.
4. CARMICHAEL, L., HOGAN, H. P., and WALTER, A. A. "An Experimental Study of the Effects of Language on the Reproduction of Visually Perceived Form," *Journal of Experimental Psychology,* Vol. XV, No. 1.
5. COLLIER, JOHN. "Wet Saturday." In *Short Stories from the "New Yorker,"* pp. 171–78. London, 1951.
6. DEUTSCH, M. *Conditions Affecting Cooperation,* pp. 113–14. Naval Research Contract, NONR-285 (10). Research Center for Human Relations, New York University, February, 1957.
7. DUFFUS, R. L. "The Function of the Racketeer," *New Republic,* March 27, 1929, pp. 166–68.
8. DULLES, J. F. "Challenge and Response in U.S. Policy," *Foreign Affairs,* October, 1957.
9. ———. Transcript of the Remarks by Secretary of State Dulles at His News Conference, *The New York Times,* August 29, 1954, p. 4.
10. FLOOD, M. M. *Some Experimental Games.* The RAND Corporation, Research Memorandum RM-789. Santa Monica, June 20, 1952.
11. GOFFMAN, ERVING. "On Face-Work," *Psychiatry: Journal for the Study of Interpersonal Processes,* Vol. XVIII, No. 3. 1955.
12. GRODZINS, M. "Metropolitan Segregation," *Scientific American,* Vol. CXCVII, No. 4 (October, 1957).
13. HARSANYI, J. "Approaches to the Bargaining Problem before and after the Theory of Games: A Critical Discussion of Zeuthen's, Hick's, and Nash's Theories," *Econometrica,* Vol. XXIV (1956).
14. HEISE, G. A., and MILLER, G. A. "Problem Solving by Small Groups Using Various Communication Nets." In P. A. HARE, E. F. BORGATTA, and R. F. BATES, *Small Groups.* New York: A. A. Knopf, 1955.
15. KAPLAN, M. A. *System and Process in International Politics.* New York: John Wiley & Sons, 1957.
16. KAYSEN, C. "A Revolution in Economic Theory?" *Review of Economic Studies,* Vol. XIV, No. 35 (1946–47).
17. KEYNES, J. M. *The General Theory of Employment, Interest and Money.* New York: Harcourt, Brace & Co., 1936.
18. KISSINGER, H. A. *Nuclear Weapons and Foreign Policy.* New York: Harper & Bros., 1957.
19. KOFFKA, K. *Principles of Gestalt Psychology.* London: Routledge & K. Paul, 1955.
20. LEAVITT, H. J., and MUELLER, R. A. H. "Some Effects of Feed-Back on Communication." In P. A. HARE, E. F. BORGATTA, and R. F. BATES, *Small Groups.* New York: A. A. Knopf, 1955.
21. LUCE, R. D., and RAIFFA, H. *Games and Decisions.* New York: John Wiley & Sons, 1957.
22. MARSCHAK, J. "Elements for a Theory of Teams," *Cowles Foundation Papers,* No. 94 (New Series).
23. MARSCHAK, J., and RADNER, R. "Structural

[44] Alex Bavelas (2, p. 493) has described an experiment in pure co-ordination in which each of five separated players must pass geometric pieces among themselves until they reach a distribution of the pieces that permits the formation of five separate squares. The pieces are so cut that many "wrong" squares can be formed, i.e., squares that use a combination of pieces that makes it impossible for four more squares to be formed with the remaining pieces. He is interested in what happens when these deceptive "successes" occur. "For an individual who has completed a square it is understandably difficult to tear it apart. The ease with which he can take a course of action 'away from the goal' should depend to some extent upon his perception of the total situation. In this regard the pattern of communication should have well-defined effects. . . . Preliminary runs . . . have revealed . . . that the binding forces against restructuring are very great, and that, with any considerable amount of communication restriction, a solution is improbable."

and Operational Communication Problems in Teams," *Cowles Foundation Discussion Papers, Economics,* No. 2076.

24. MARSCHAK, J. "Toward an Economic Theory of Organization and Information." *Cowles Foundation Papers,* No. 95 (New Series).
25. MOORE, O. K., and BERKOWITZ, M. I. *Game Theory and Social Interaction.* Office of Naval Research, Technical Report, Contract No. SAR/NONR-609 (16). New Haven, November, 1956.
26. NASH, J. F. "The Bargaining Problem," *Econometrica,* Vol. XVIII, No. 2 (1950).
27. ——. "Two Person Co-operative Games," *Econometrica,* Vol. XXI, No. 1 (1953).
28. "Japan Debating Atomic 'Suicide,'" *The New York Times,* March 5, 1957, p 16.
29. "Rail Strikers Sit in Tracks," *The New York Times,* May 13, 1957, pp. 14L f.
30. SCHELLING, T. C. "An Essay on Bargaining," *American Economic Review,* Vol. XLVI, No. 3 (June, 1956).
31. ——. "Bargaining, Communication, and Limited War," *Journal of Conflict Resolution,* Vol. I, No. 1 (March, 1957).
32. ——. *For the Abandonment of Symmetry in the Theory of Co-operative Games.* The RAND Corporation, Paper P-1386. Santa Monica, 1958.
33. ——. *Reinterpretation of the Solution Concept for "Non-Co-operative" Games.* The RAND Corporation, Paper P-1385. Santa Monica, 1958.
34. ——. *The Reciprocal Fear of Surprise Attack.* The RAND Corporation, Paper P-1342. Santa Monica, 1958.
35. SHERIF, M. "The Development of Social Norms." In R. W. O'BRIEN, *Readings in General Sociology.* Palo Alto, Calif., 1947.
36. SUTHERLAND, E. H. *The Professional Thief.* Chicago: University of Chicago Press, 1954.
37. SZILARD, L. "Disarmament and the Problem of Peace," *Bulletin of the Atomic Scientists,* Vol. II, No. 8 (October, 1955).
38. VON NEUMANN, J., and MORGENSTERN, O. *Theory of Games and Economic Behavior.* Princeton, N.J.: Princeton University Press, 1953.