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, that is, 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 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.[15] 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, and so forth, 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.[16]

[15] This has been evident in preliminary experiments with such a game.

[16] 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.)

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 identified 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 relations.

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 nondestructive, 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.[17]

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

[17] 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. (This point is developed at length in Appendix A.)

clues and signals the game provides are relevant not merely to the study of how players actually do behave in a nonzero-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 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.[18]

[18] It should be added that the concept of the intrinsic magnetism or focusing quality of particular outcomes in a bargaining situation or in a pure coordination 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 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 where very familiar but less simple, figures were not Koffka 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" (K Koffka, *Principles of Gestalt Psychology* [London, 1955])

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 nonzero-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

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, and so forth; 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.) [19]

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 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

[19] The following observation, quoted by Koffka, 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."

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; that is, 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 coordination 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 of where expectations should converge in the rest of the game, seems to be an area in which experimental psychology can contribute to game theory.

The Empirical Relevance of Mathematical Foci. We must avoid assuming that everything the analyst can perceive is perceived by the participants in a game, or that whatever exerts power of suggestion on the analyst does so on the participant in a game. In particular, game characteristics that are relevant to sophisticated mathematical solutions (except when the same solution can also be reached by an alternative, less sophisticated route) might not have this power of focusing expectations and influencing the outcome. They might have it only if the players perceived each other to be mathematicians. This may be 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 other suggestive and connotative details, can serve to focus the expectations of certain participants on certain solutions. If two players are themselves mathematical game theorists, they may mutually perceive and be powerfully affected by potential solutions that have compelling mathematical properties. Each may transcend, and know that the other will transcend, various adventitious details that, to

nonmathematician game players, might be more relevant to the focusing of expectations than some of the quantitative properties of the game.

(In many cases these mathematical properties would be a uniqueness or symmetry that would have nonmathematical definitions and nonmathematical appeal, too, or would happen to coincide with qualitatively distinguishable points that could be rationalized in an equally compelling nonmathematical way.)

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.

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. (This topic is pursued further in Appendix B.)

One may or may not agree with any particular hypothesis 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 coordinated choice, should be analytically characterized by the notion of converging expectations.

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?[20]

[20] 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

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 everbody 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 options, if only because such facts are inherently unknowable or incommunicable.

Von Neumann and Morgenstern illustrated their *solution* concept for the nonzero-sum game with the example of a seller, A, prepared to sell his house for any price above 10, and two buyers, B and C, prepared to pay up to 15 and 25, respectively.[21] (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.

---

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, "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." "On Face-Work," *Psychiatry: Journal for the Study of Interpersonal Processes*, 18:224 (1955).

[21] J. Von Neumann and O. Morgenstern, *Theory of Games and Economic Behavior* (Princeton, 1953), pp. 564ff.

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.

5

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 nonzero-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 may have defined away some of the essential ingredients of typical nonzero-sum games. Preoccupied with the solution to *the* nonzero-sum game, game theory has not done justice to some typical game situations or game models and to the "moves" that are peculiar to nonzero-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 enforce-

ment 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;[1] 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.[2] 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 twenty 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

[1] Compare C. W. Sherwin, "Securing Peace Through Military Technology," *Bulletin of the Atomic Scientists*, 12:159–164 (May 1956).

[2] Compare Herman Kahn and Erwin Mann, "Game Theory," The RAND Corporation, Paper P-1166 (Santa Monica, 1957), pp. 55ff. The authors work out a number of problems involving dynamite, detonators, and deterrence

together, he can deter agreement. If the twenty 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 fourteen of the twenty can overpower the remaining six and force them to advance, they can demonstrate that they could overwhelm the man; if so, the threat succeeds and the gunman surrenders, and even the six "expendables" gain through their own inability to avoid jeopardy. If the twenty 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 nineteen 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 chapter 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

An example of a standard "move" is the commitment, analyzed at some length in Chapter 3. If the institutional environment makes it possible for a potential buyer to make a single "final" offer subject to extreme penalty in the event he should amend the offer — to commit himself — there remains but a single, well-determined decision for the seller: to sell at the price proposed or to forego the sale. The possibility of commitment converts an indeterminate bargaining situation into a two-move game; one

player assumes a commitment, and the other makes a final decision. The game has become determinate.[3]

This particular move, analyzed at length in Chapter 3, is mentioned here only as a particularly simple illustration of a typical move. As noted in Chapter 3, 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. Furthermore, 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.)

But, 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

[3] In the real estate example of Von Neumann and Morgenstern referred to earlier (p. 116) 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 (that is, 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 even greater than was B's originally, if the commitment is demonstrable while subjective valuations are not.

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 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." [4] 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 contingency being governed by the second party's behavior. Like the ordinary commitment, the threat is a surrender of choice, a renunciation of alternatives, that makes one worse off than he need be in the event 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. Like an ordinary commitment, a threat can constrain the other player only insofar as it carries to the other player at least some appearance of obligation; 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.[5]

[4] *Games and Decisions*, p. 75.

[5] In ordinary language, "threat" is often used also for the case in which one merely points out to an adversary, or reminds him, that one would take action painful to the adversary if the latter fails to comply, it being clear that one would have incentive to do so. To "threaten" to call the police on a trespasser is of this sort, the threat to shoot him is not. But it seems better to use a different word for these cases — I suggest "warning" rather than "threat" — because the "threat" either is superfluous, and does not constitute a move, or it conveys true information and relates to situations with an information struc-

The threat differs from the ordinary commitment, however, in that it makes one's course of action *conditional* on what the other player does. While the commitment fixes one's course 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 therefore 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 player 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

---

ture and communication structure worth keeping distinct. In this latter case it is a mutually beneficial move, precluding a jointly undesired outcome by improving the second party's understanding The main point of analytical similarity, between this "warning" case and that of the "threat," is in the possible difficulty of conveying true information credibly, of conveying *evidence* for the assertion that one would have, ex post, incentive for doing as one warns he will As a matter of fact, if a threat is of such nature (as it often is) that the act of commitment is not contained in the act of communicating it — if the commitment precedes the conveyance of the threat, with evidence for believing it, to the threatened party — the first act in the process of threatening changes the "true" incentive structure, and the second is, in effect, a "warning"

Luce and Raiffa, 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." [6]

[6] Pp. 110–11, 119–20, 143–44. Morton A. Kaplan, 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." (See his *System and Process in International Politics* [New York, 1957].) Luce and Raiffa may partly be led to their view that only one of the 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 — that is, the legal equivalent of "first move" in advance of his partner's threat. The "threat" tactic of J. F. Nash, 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 nonagreement. This is the kind of threat assumed by Luce and Raiffa (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

But the issue is both simpler and more precise than that. Consider the left-hand matrix in Fig. 9, 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

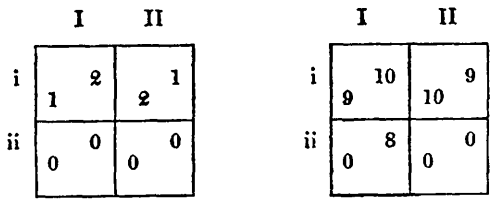

FIG. 9

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 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

---

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 nonagreement but with the public not concerned with what the agreement provides as long as some agreement is reached.

pair to choose from — ii,I versus i,II — there is no doubt what Column will do. If I threaten to blow my 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.[7]

The issue here is in whether or not we admit that the game has "moves," that is, 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 nonfulfilment 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

[7] Edward Banfield showed me this irresistible quotation about the Bhâts and Charáns of the west of India, revered as bards. "In Guzerát they carry large sums in bullion, through tracts where a strong escort would be insufficient to protect it. They are also guarantees of all agreements of chiefs among themselves, and even with the government.

"Their power is derived from the sanctity of their character and their desperate resolution. If a man carrying treasure is approached, he announces that he will commit trága, as it is called: or if an engagement is not complied with, he issues the same threat unless it is fulfilled. If he is not attended to, he proceeds to gash his limbs with a dagger, which, if all other means fail, he will plunge into his heart; or he will first strike off the head of his child; or different guarantees to the agreement will cast lots who is to be first beheaded by his companions. The disgrace of these proceedings, and the fear of having a bard's blood on their head, generally reduce the most obstinate to reason. Their fidelity is exemplary, and they never hesitate to sacrifice their lives to keep up an ascendency on which the importance of their cast depends" (The Hon. Mountstuart Elphinstone, *History of India* [ed. 7; London, 1889], p. 211).

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 nonfulfilment 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, 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." [8]

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 do. This concept can be loosened by supposing that the penalty is of finite size and not necessarily so large as to be controlling in all cases. In Fig. 10 Column will

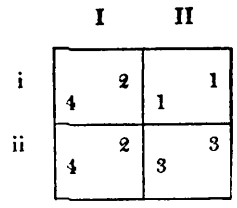

FIG. 10

[8] 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 — the law.

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 Fig. 11 without threats, the solution

	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. 11

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.[9] Column would threaten I against iii, Row would threaten i against

[9] 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 chose 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 (that is, for commitments to the second party) are generally of a quite different nature from those for unilateral commitments (that is, 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, that is, to be able to comply.

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 noncompliance 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, 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* 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.[10] 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

[10] Situations of this sort are explored in Chapters 7 and 9.

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 noncompliance 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 Fig. 12, if choices

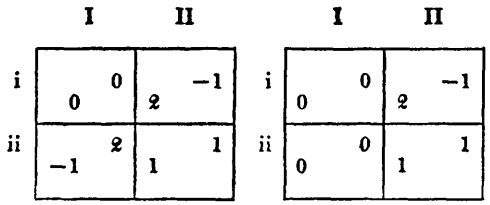

FIG. 12

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 *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.[11] 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.

The exact definition of a promise — for example, 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 Fig. 12. But Fig. 13 shows a situation in which

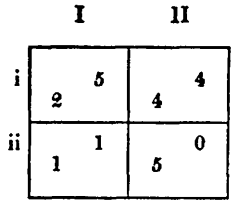

FIG. 13

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 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 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

[11] This notion is celebrated in "Wet Saturday," by John Collier, 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. (*Short Stories from the "New Yorker"* [London, 1951], pp. 171–178.)

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.[12]

A better definition, perhaps, would make the promise a commitment that is controlled by the second party, that is, 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 noncompliance 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 noncompliance 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 relation between parties and the recognition by each that what he might

[12] 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.

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 coordination 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 into their agreement certain elements whose sole purpose is to create severe jeopardy for noncoordination. 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.[13]

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

[13] The precise definition of hostages is a little difficult. They seem to be as pertinent to threats as to promises: the 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 or "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.

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 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.) [14]

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.[15] 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 nonperformance. 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

[14] This concept is developed at length in Chapter 10.

[15] 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, "The Function of the Racketeer," *New Republic* (March 27, 1929), pp. 166–68.

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 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.[16]

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 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

[16] J. F. Dulles, "Challenge and Response in U. S. Policy," *Foreign Affairs* (October, 1957). Very similar language is used by Dean Acheson (*Power and Diplomacy* [Cambridge, Mass., 1958], 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."

nicely accomplished by Lieutenant Colonel (then Major) Stevenson B. Canyon, U.S.A.F., in using his aircraft to protect a Chinese Nationalist 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 to avoid the flames. 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 nonresistance." According to *The New York Times*, "Striking railway workers sat down on the tracks at more than 300 stations in Japan today, halting 48 passenger and 144 freight trains." [17]

A more dramatic instance, also Japanese, was reported in the same paper: "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." [18]

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

[17] "Rail Strikers Sit in Tracks," *The New York Times* (May 13, 1957), pp. 14L f. 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.

[18] "Japan Debating Atomic 'Suicide,'" *The New York Times* (March 5, 1957), p. 16.

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*. Angelo, acting in place of the 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 his life. Angelo, finding the sister attractive, proposes a dishonorable bargain; the sister declines, Angelo 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. Angelo'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 Angelo — 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: "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." [19]

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,

[19] E. H. Sutherland, *The Professional Thief* (Chicago, 1954), p. 126.

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 Fig. 14 — disregarding the numbers in parentheses — if Row has second move, he loses in the lower right-hand

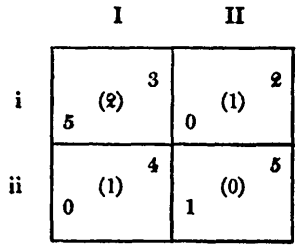

FIG. 14

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. The payoffs of the latter are such that with second move 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, counter-offers, and so forth — 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 coordination 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. The principle is beautifully illustrated by the daylight-saving-time controversy; a majority that wants to do everything an hour earlier just cannot organize to do it unless it gets legislative control of the clock. And when it does, a well-organized minority that opposed the change is usually quite unable to offset the change in clock time by any organized effort to change the nominal hour at which it gets up, eats, and does business.

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.[20]

[20] 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

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 coordination, 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 chapter, 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 struc-

---

to be common traits or where a reputation for them is considered of much higher value.

ture 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 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 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.[21]

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.

[21] L. Szilard, "Disarmament and the Problem of Peace," *Bulletin of the Atomic Scientists*, 2:297–307 (October, 1955).

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 noncompliance; 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.[22]

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

[22] What might be called the "legal status" of communication is nicely developed by Goffman: "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" (pp. 224, 226).

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 — that is, 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 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.[23] 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

[23] Daniel Ellsberg, some of whose work in the field of strategy was contained in the lectures mentioned in Chapter 1, independently arrived at precisely this formulation of the threat or commitment, namely, as a selective reduction of some of one's own payoffs in the strategy matrix.

the penalty from one's own payoffs in all cells that do not correspond to the strategy so selected.[24]

Specifically, in Fig. 15 *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 domi-

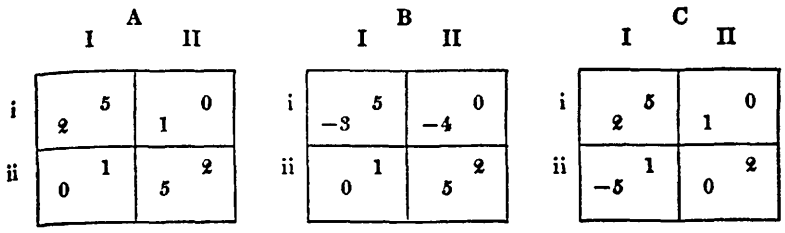

FIG. 15

nant strategy, that is, a strategy that he would follow no matter which column the other player selects. The result would be the modified matrix shown in Fig. 15 *B*. (Committing himself to i

[24] 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. (Further generalization would include randomized strategies; these are introduced in Chapter 7.)

	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. 15 *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 Fig. 15 *A*, but also the strategies of *commit, threaten, promise,* and so forth? 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 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 noncommitment.

If we put these strategies into matrix form, we get Fig. 16. The 12 × 8 matrix of Fig. 16 represents the tacit ("noncooperative") 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 specifica-

		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. 16

tion of strategies in the enlarged version of the game; they *are* strategies of response or adaptation.

This is brought out in the labeling of Fig. 16. 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 "o" 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 o-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 o, 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 i, then choose row i no matter what Column does." Knowing the payoffs in the original game, Fig. 15 *A*, the players can identify the payoffs in the enlarged game of Fig. 16. 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 (that is, given one particular strategy for the enlarged game). To determine what instructions to give, Row and Column consider the matrix in Fig. 16; 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 Fig. 16, 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 Fig. 15, in which he has second move.)

How do we characterize the cells, or pairs of strategies, that represent the "solution" in Fig. 16? They constitute a solution of the kind that has been called a *solution in the complete weak sense.*[25] It can be arrived at, within the framework of the enlarged matrix, 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 of the original 2×2 matrix, Fig. 15 *A*, as Column can expect him to do, it makes no differ-

[25] Compare Luce and Raiffa, pp. 106–09.

ence what instructions Column gives his agent regarding the two contingencies that did not arise.) [26]

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 enforcements 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

[26] It is worth noting that the order in which we discard the rows and columns that are eligible for discard can affect the form of the "solution." In the procedure outlined in the text, we first discarded all rows but the third and tenth; we then observed that columns I, III, V, and VII, were eligible for discard, and discarded them; at that stage, row iii was seen to be dominated, and it was discarded; and we were left with row x intersected by four columns that yielded identical payoffs in that row. But we might have noted, as we discarded the four columns, that two more columns could also be discarded at that stage, namely columns VI and VIII, which show inferior payoffs to Column, *in row iii*, than columns II and IV. In other words, at that point in the process, row iii and columns VI and VIII were all eligible for discard; but if we arbitrarily choose first to eliminate row iii and *then* proceed to the columns, the two columns in question are no longer dominated. Thus, in a sense, the contents of our "solution" depend on an arbitrary choice of procedure; whether we are left with two cells with identical payoffs, however, or four cells with identical payoffs, depends on that arbitrary choice. The payoffs, however, are the same in either case. The rationale might be that at some stage

that make up the move-structure of games; too abstract a model will miss them.[27]

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 threat-

---

Column sees that he needn't reason any further, that Row has a clearly determined choice that makes it inconsequential whether Column further narrows his decision, but that the exact point at which he perceives this, and what columns are left uneliminated when he does perceive it, depends to some extent on which of several alternative routes he pursues in his reasoning process. (If there were communication costs in narrowing his choice of strategy, Column might prefer to choose strategy 2–II only, leaving unspecified what choice would correspond to Row's strategy 0 or 1. If, to take a contrary case, there are risks that Row's strategy will be erroneously recorded or communicated, or unintelligently chosen, Column reduces his risks by specifying 0–I as well. In the latter case he, in effect, treats row iii as not wholly unlikely in spite of its domination by row x. And if, to take the matter further, he suspects that the referee has a tendency to hear "row v" when other rows are actually chosen, he may further narrow his choice to 0–I, 1–I, 2–II, the "solution" being the intersection of row x and Column II, since the intersection of v and IV is inferior to that of v and II and gives him grounds for this further refinement of his choice. In general, by attaching risks of error of various sort, or differential costs of different ways to specify a strategy, a rather richer problem is formed, and one that can lead to different conclusions. The problems treated in Chapters 7 and 9, involving certain forms of random behavior, error, or misinformation, can produce this kind of result.)

[27] Incidentally, casting a particular game into supergame matrix form is generally not a feasible technique of analysis; the number of rows and columns (that is, 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.

ened 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 diagrammatic form, was illustrated in the final paragraph of Chapter 2, and referred to as an abstract example 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 while facing an enemy 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.[28]

[28] 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, 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

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 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 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." [29] 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 every-

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

[29] Transcript of the Remarks by Secretary of State Dulles at His News Conference, *The New York Times* (August 29, 1956), p. 4.

one 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.) [30]

"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 striking, 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.

[30] That one's position can be painfully weakened by new legal powers is poignantly suggested by one of the arguments raised against legalizing euthanasia, granting hopeless incurables the right to authorize their own removal: "What . . . would be the effect on old people with incurable infirmities who are already suspicious that those around them want to get rid of them?" (John Beavan, "The Patient's Right to Live—and Die," *The New York Times Magazine*, August 9, 1959, pp. 14, 21–22.)

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

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. And, similarly, in a zero-sum game between rational players with full information it can never be an advantage to move first (to play the "minorant game" in the language of von Neumann and Morgenstern); in the mixed game it certainly can.

6

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 nondestructive 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 coordination, we have to recognize that the *kinds* of things 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 nonzero-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.[1]

[1] A good laboratory example of the communication-perception part of game strategy is the experiment reported by M. M. Flood, who presented his players with a 2 × 2 nonzero-sum matrix for 100 consecutive tacit plays. The special property of the matrix is that the players can win only by cooperating on a particular cell on each play, but to distribute the winnings for the 100-play sequence they must cooperate 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 uncoordinated choice is a lost chance to make some money. M. M. Flood, "Some Experimental Games," *Management Science*, 5:5–26 (October, 1958).

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 psycholo-

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 communication. But in a nonzero-sum game that involves any initial uncertainty over which among the possible outcomes are in fact efficient and any need for coordinated 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 cooperative 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 empirically 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

---

gists 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.)

that was made nonzero-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, that is, a mutually nondestructive 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.[2]

[2] 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

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 center 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, that is, "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?

---

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.

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 cooperatively, 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 nonprice competition in a particular industry), it seems likely that the empirical part of game theory will include experimental work like that of Muzafer Sherif. 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, that is, 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.[3] In an almost unconsciously cooperative way, adversaries must reach a mutually recognized definition of what constitutes an innovation, a challenging or assertive move, or a cooperative gesture, and they must develop some common norm regarding the kind of retaliation that fits the crime when a breach of the rules occurs.[4]

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 particular 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

[3] A splendid example of the creation of norms in practice — and one that suggests that the process is susceptible of analysis — was the rather general 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.

[4] 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 Deutsch is pertinent. He let pairs of players play nonzero-sum games (in matrix form) tacitly for a sequence of two plays, the game providing both a "cooperative" and an "uncooperative" choice. Those who played uncooperatively against a cooperative partner had an opportunity, on the second play, to respond to the implicit offer of cooperation. 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." See Morton Deutsch, *Conditions Affecting Cooperation*, Research Center for Human Relations, New York University, 1957. (An article based on this monograph, not including the point quoted here, entitled "Trust and Suspicion," appeared in *The Journal of Conflict Resolution*, 2:265–279 [December 1958].)

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 affected 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 "nonauthoritative" way, a background story that suggestively indicates some other hypothetical starting point.[5]

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 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

[5] The income-tax questions described in Chapter 3 (pp. 62–65) indicate the force of this power of suggestion.

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 (that is, 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.[6]

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 coordination 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

[6] "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" (Henry A. Kissinger, *Nuclear Weapons and Foreign Policy* [New York, 1957]).

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 communication systems often lends itself to systematic experimental study.[7]

[7] Alex Bavelas has described an experiment in pure coordination 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, that is, 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*" ("*Communication Patterns in Task-oriented Groups*," in D. Cartwright and A. F. Zander, *Group Dynamics* [Evanston, 1953], p. 493). Some very suggestive experimental work, especially on "the biased perception of what is equable," is reported by Charles E. Osgood, "Suggestions for Winning the Real War with Communism," *Journal of Conflict Resolution*, 3:304–05 (December, 1959).

PART III

STRATEGY WITH A RANDOM INGREDIENT

7

RANDOMIZATION OF PROMISES AND THREATS

In the theory of games of pure conflict (zero-sum games) randomized strategies play a central role. It may be no exaggeration to say that the potentialities of randomized behavior account for most of the interest in game theory during the past one and one-half decades.[1] The essence of randomization in a two-person zero-sum game is to preclude the adversary's gaining intelligence about one's own mode of play — to prevent his deductive anticipation of how one may make up one's own mind, and to protect oneself from tell-tale regularities of behavior that an adversary might discern or from inadvertent bias in one's choice that an adversary might anticipate. In the games that mix conflict with common interest, however, randomization plays no such central role, and the role it does play is rather different.[2]

[1] John von Neumann, speaking of "the fundamental theorem on the existence of good strategies," namely the theorem that all zero-sum games with a finite number of pure strategies have a minimax-maximin equilibrium pair ("solution") if mixed strategies are allowed, said, "As far as I can see, there could be no theory of games on these bases without that theorem. . . . Throughout the period in question I thought there was nothing worth publishing until the 'minimax theorem' was proved" ("Communication on the Borel Notes," *Econometrica*, 21:124–125 [January 1953]).

[2] One can, instead, interpret mixed strategies in zero-sum games as a means of introducing continuity of strategies into a discrete-strategy game that has no pure-strategy saddle point, thereby converting it into a game that does have a saddle point. In this interpretation the role of mixed strategies in zero-sum games is not so different from their role in the nonzero-sum games. One can flip a coin to keep an opponent from guessing with confidence whether it will come up heads or tails; or one may flip a coin to "average" heads and tails, to create (in an expected-value sense) a strategy halfway between heads and tails. Both interpretations are useful. If the second is somewhat more sophisticated, the first may better catch the spirit of the problem as it presents itself to a game player. And the first reminds us that the problem, even with

Randomization in the theory of these ("nonzero-sum") games is not mainly concerned with preventing one's strategy from being anticipated. In these games, as noted earlier, one is often more concerned with making the other player anticipate one's mode of play, and anticipate it correctly, than with disguising one's strategy.

There may of course be zero-sum components embedded in a larger game. In limited war one may be concerned to communicate rather than to disguise the limits that one proposes to observe, but within those limits may sortie his aircraft in a randomized way to minimize the enemy's tactical intelligence.[3] Again, information samples may be exchanged, or agreements enforced on a sample basis, where neither party can afford to yield the other full knowledge. Arms-control agreements, for example, might have to be monitored by a sampling technique that yielded each side enough knowledge about the enemy's forces to reveal compliance or noncompliance without yielding so much that the possibility of successful surprise attack on those forces were greatly enhanced.

But the main role of randomization in the traditional literature on nonzero-sum games is a different one. It has been a device to make indivisible objects divisible, or incommensurate objects homogeneous. Their "expected values" are divisible by lottery when the objects themselves are not. We flip coins to see who

---

randomization, is still to prevent the opponent's anticipation of our actual strategy choice, and that the machinery of choice, the procedures for recording and communicating a choice, and any advance preparations required by the outcome of the random process, must remain inaccessible to his intelligence system.

[3] In particular cases there may be a tantalizing dilemma inherent in a choice of secrecy or revelation. If in order to prove that one is committed to a threat, or that one is in fact capable of fulfilling the threat, one must display evidence of the commitment or the capability to the other party, the evidence may be of a kind that necessarily yields information helpful to the second party in combatting the threat. To prove to an enemy that one has a potent weapon that can overcome his defenses we might have to demonstrate the weapon or some aspect of it, or provide technical knowledge to prove the weapon feasible; to do so may aid him greatly in preparing a defense against it. If, to prove we would fight a local war in an ambiguous area, it were necessary to station troops there ahead of time, the enemy would have the advantage of knowing their exact location rather than having to be prepared in all directions.

gets the object, and play "double or nothing" when we cannot make change. We can divide the obligation of citizenship equally by selecting draftees through a lottery, when we want a fraction of the eligibles for a long period of service rather than all of them for a short one.

In this role, randomization is evidently relevant to promises. If the only favors available to be promised are larger than necessary and not divisible, a lottery that offers a specified probability of the favor's being granted can scale down the expected value of the promise and reduce the cost to the person making it. An offer to help a person on a large scale in a contingency is somewhat equivalent to offering the certainty of smaller help. (There may be the additional advantage that the contingency is correlated with his need.)

But in this respect a promise is different from a threat. The difference is that a promise is costly when it succeeds, and a threat is costly when it fails. A successful threat is one that is not carried out. If I promise more than I need to as an inducement, and the promise succeeds, I pay more than I needed to. But a threat that is "too big" is likely to be superfluous rather than costly. If I threaten to blow us both to bits when it would have been sufficient to threaten our discomfort, you'll likely still comply; since I have neither to discomfort us nor to kill us, the error costs nothing. If all I had was a grenade to explode in our midst and wished for tear gas instead, I might scale down the grenade to the "size" of a tear-gas bomb by threatening an appropriate percentage chance that the bomb would go off, killing us both, if you failed to comply. But the need to do this is not as clear as in the case of a promise, where any excess in the value promised is so much loss.

The size of the threat can be a problem if it costs something to be equipped to *make* a threat and if bigger threats cost more to make than small ones. If a threat of tear gas is enough, so that I do not need to threaten explosion, and if tear-gas bombs are cheaper than explosive ones, and if I have to display the bomb to make the threat persuasive, it is better to threaten with the cheaper tear gas. But grenades may be cheaper, and then the incentive goes the other way. For many interesting threats

the greatest cost is the risk of having to carry it out, and the more ordinary "cost" is not a controlling factor.

THE RISK OF FAILURE

The risk of *failure*, however, does give an incentive to choose moderate rather than excessive threats. If the only threat that can be made is some horrendous act, one may be tempted to scale it down by attaching it to a lottery device — by threatening some *specified probability* that it will be carried out unless compliance is forthcoming, not by committing oneself to the certainty that the jointly painful punishment would be administered.

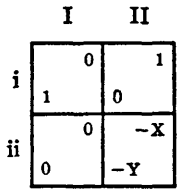

FIG. 17

To illustrate, consider the matrix in Fig. 17, in which Column has first choice, followed by Row, but in which Row has the option of making a prior threat to constrain Column's choice. (Interpret X and Y as positive numbers.) On one condition, Row's strategy is clearly to threaten row ii if Column chooses column II. If he makes no threat, Column chooses II knowing that Row will then choose i. Given the threat — and assuming that Row is committed to it and that Column knows it — the choice of II yields unattractive outcomes for both of them, and Column can be expected to choose I.

The condition is that Row be quite sure that nothing will go wrong! Maybe he completely misjudges Column's payoffs. Maybe this particular adversary is drawn from a universe in which nearly everyone, but not quite everyone, has preferences as indicated in the matrix, and a few deviants have a radically different preference system and prefer the lower right cell to the upper left one. Alternatively, Row may get himself committed to his threat but

fail to communicate it convincingly to Column, so that Column mistakenly ignores the threat, condemning them both to the lower right-hand cell. Again, Column himself may have arranged a prior commitment through his own choice of II, and failed to communicate it accurately to Row in time for Row to take this into account, or Column may have suffered a disability unknown to Row that eliminates the possibility of I; in that case, Row's own commitment will only guarantee the worst outcome for both players. Whatever the reasons for failure, there is perhaps some probability that the threat will fail. If we take it into account we may have a reason for Row to wish that the "punitive" payoffs in the lower right-hand cell were not quite as unattractive as they are.

If Row is confined to "pure" strategies — if he must specify his threat or commitment without reference to error or chance — he can do nothing but wish that the numbers in the lower right-hand cell were not so unattractive. But if he can randomize his threat he can in fact "scale it down" to reduce somewhat the high cost of failure. If, for example, he can commit himself not to a choice of row ii in the event that column II is chosen, but to a 50-50 chance between i and ii in that event, he may still hope to frighten Column into a choice of I while reducing the seriousness of the risk of failure.

We can be more specific. Let P stand for the probability that the threat will fail for any reason whatsoever. (For our present purpose this is an "autonomous" probability, independent of Row's strategy.) Let Row now threaten to choose ii with probability equal to π, in the event Column chooses II. In other words, if Column fails to comply there is a probability of π that Row will choose ii to their mutual discomfort, and of $(1 - \pi)$ that he will choose i to their mutual relief. What value of π should Row choose?

First, how large does π have to be to make the threat effective at all, that is, to make it effective assuming that it does *not* fail for any of the autonomous reasons involved in P? This is a question of Column's choice when he is confronted with the risk π. If Column chooses I he gets 0. If he chooses II his expectation is a weighted average of 1 and $-X$, with weights of $(1 - \pi)$ and π

respectively. If this average is less than 0, he is motivated to choose I — subject to the autonomous probability, P, that for one reason or another he will choose II in spite of his apparent motivation toward I. The condition for an effective threat is thus [4]

$$0 > (1 - \pi) - \pi X,$$
$$\pi > \frac{1}{1 + X}.$$

Second, assume that any threat with π above the floor established by the preceding formula will succeed or fail with probabilities $(1 - P)$ and P respectively. If the threat succeeds, Row's payoff is $+1$. If it fails, his expectation is a weighted average of 0 and $-Y$, the weights being $(1 - \pi)$ and π respectively. The expected value of the outcome, then, when the threat is large enough to be effective at all, is given by

$$(1 - P) + P(0 - \pi Y) = 1 - P - P\pi Y.$$

This value is evidently higher, the lower is the value of π. Row should therefore arrange the lowest value of π that he can that meets the first condition. For a threat to be worthwhile at all — to have an expected value greater than zero, which is what Row can expect from this particular matrix if he makes no threat — a value of π must be arranged that meets the condition

$$1 - P - P\pi Y > 0$$

or

$$\frac{1 - P}{P} \cdot \frac{1}{Y} > \pi.$$

Thus the effective range for π in this example is given by

$$\frac{1 - P}{P} \cdot \frac{1}{Y} > \pi > \frac{1}{1 + X}.$$

And there is no threat at all worth making if there is no room between these two limits, if

[4] Since the analysis depends only on comparisons of the *differences* between absolute valuations of the payoffs for the two players separately, no violence is done by adopting, for each player, a scale of measurement that sets his preferred payoff equal to $+1$ and his next preferred payoff to 0. The full interpretation, then, of the expression $1/(1 + X)$, is: the ratio of (1) the difference between Column's upper right and upper left payoffs, to (2) the sum

$$\frac{1-P}{PY} < \frac{1}{1+X}$$

or

$$\frac{P}{1-P} > \frac{X+1}{Y}.$$

Only a "fractional" threat — a threat with π less than 1 — is worth making if:

$$\frac{1-P}{PY} < 1$$

or

$$\frac{P}{1-P} > \frac{1}{Y}.$$

Here is a case, then, in which the fractional threat is superior to the certainty threat, and in which the latter could be not worth making at all while the former were. The argument hinges on the risk of failure, a risk that has been assumed independent of the size of π itself. This is a somewhat special assumption. If we interpret P as the probability that we have misjudged our adversary and exaggerate his preference for avoiding the lower right cell, our assumption implies a bimodal distribution of payoffs in the population. It implies that we have either a man whose payoffs are adequately represented by the numbers in our matrix, or a man whose payoffs are so different that no relevant threat — within the range of values up to $\pi = 1$ — will dissuade him. If instead we supposed that the ratio of column payoffs in the upper and lower right-hand cells showed a bell-shaped frequency distribution within the population, and that our particular adversary had been drawn at random, the probability that our threat would succeed would vary directly with

of the differences between (a) his upper right and upper left payoff and (b) his lower right and upper left payoffs. The simplicity of the formulae thus reflects advantage already taken of this scaling convenience. It takes only one parameter to characterize the relevant relations among three valuations. (In a later problem that involves the lower left cell, all four payoffs are relevant and a second parameter would be required. That case, however, can be further simplified if the lower left payoff can be taken equal to one of the others and still illustrate the point; we get less complete knowledge but more 0's and 1's that way.) On the interpretation of these numbers see A. A. Alchian, "The Meaning of Utility Measurement," *American Economic Review*, 43: 26–50 (March, 1953), or Luce and Raiffa, pp. 12–38.

the value of π itself. The probability that a burglar drawn at random from the universe of burglars will be deterred by some specified probability of apprehension and conviction presumably varies directly with the latter probability; the simple model analyzed above treats burglars as divisible into two classes — those, let us say, who steal for money and are certainly deterred in accordance with the numbers of the matrix, and those who steal for fun and are beyond reach of any threat of the magnitude entered in the lower right-hand cell. On the other hand, if our probability of failure reflected, say, a breakdown of communication with the adversary, there might be better reason for supposing the probability of failure to be independent of the particular threat being communicated.

It is interesting to notice that attaching a probability of fulfillment to our threat is, in the above model, substantially equivalent to scaling down the size of the threat more directly. To see this, interpret X in the lower right-hand cell as a fine that will be levied on both Row and Column, or a number of lashes with the whip or days of imprisonment that both will suffer if the threat is fulfilled. If X is the maximum number of dollars, lashes or days that Row can threaten, let π be interpreted as Row's specification of what fraction of the maximum permissible penalty is to be exacted; if π is set at 0.5, for example, both Row and Column receive exactly half their maximum punishments. If we interpret the matrix in this way, and ask what value of π provides the optimum threat from Row's point of view, we go through the same analysis and we reach the same conclusion as before, namely, π is to be as small as possible subject to a minimum value equal to $1/(1 + X)$. Thus we can interpret π either as a probability of threat fulfillment or as the scale on which the threat is to be certainly carried out. Since the two formulations come to the same thing, and we can interpret π either way, it seems fair to say that *in this case* the role of randomization is that of making divisible an otherwise too large and indivisible threat, of making possible a "smaller" threat than was otherwise available. (It should be noted though, that to reduce a threat by reducing the probability of its fulfillment reduces the expected value of the outcome proportionately for both players, while a

direct reduction in size might not be restricted to proportionate changes in value or utility for the two parties.) [5]

THE RISK OF INADVERTENT FULFILLMENT

There is another "cost" element that can motivate a reduced threat. This is the risk that one will fulfill the threat inadvertently, even if the adversary does comply with it (or would have complied if the threat hadn't gone off accidentally before he had a chance). The gun that threatens a burglar or hold-up victim may go off accidentally before he has a chance to comply. The dog that threatens to bite trespassers may bite some who do not trespass.

If a hitchhiker pulls a gun on the driver of a car and the driver threatens to kill them both unless the hitchhiker throws his gun out the window, making his threat by pressing the accelerator to the floor and creating a manifest risk of fatal accident, there is some chance that the accident will occur before the hitchhiker has a chance to comprehend the threat and comply. In this case, the risk of accidental fulfillment is an integral part of the threat. The only way one can make the threat is to start fulfilling it. Until the driver speeds up the hitchhiker has no reason to believe him; once he does speed up, there is some minimum length of time it takes the hitchhiker to comply and the driver to relax his speed. There is therefore an interval, however short it may be, that the risk is present; the risk entailed by the high speed must therefore be one that is small enough to be tolerable to the driver during this initial interval. If instead the car were definitely safe at all speeds under sixty but would certainly skid off the road at exactly sixty and there were no gradations between that carried a moderate risk of accident, the driver could have no incentive to incur a dangerous speed and the hitchhiker would know it and not respond to a verbal threat of high speed. It is the possibility of a "fractional threat," a threat that carries the risk but not the certainty of death, that gives the driver

[5] Randomization may also be integrally related to the arrangement of the threat itself, or be involved in the decision process whether the threatener wishes it or not. So the interpretation of randomization as just a means of manipulating the size of the threat is applicable only in some cases.

anything to work with; but to put it into effect he has to suffer it for some finite period.

If in situations of this kind we suppose — as is roughly true in the hitchhiker case — that the risk of inadvertent fulfillment is proportionate to the probability, π, that one will fulfill the threat if the adversary does not comply — if the watchdog's propensity to bite innocent passersby is proportionate to his proclivity to bite those who enter the premises — a formula is obtained that is not very dissimilar to the one already arrived at. Using the same matrix as before (ignoring this time the probability that a potentially effective threat may fail) and letting $a\pi$ represent the probability of inadvertent fulfillment, the minimum value of π is the same as before. The expected value of the outcome to Row, which must exceed 0 if he is to make the threat, is given by the left-hand side of the formula

$$(1 - a\pi) - a\pi Y > 0,$$

or

$$\frac{1}{a(1+Y)} > \pi > \frac{1}{1+X}.$$

The optimal threat is again one that barely exceeds the lower limit; there is an upper limit to π that may be less than 1; and, depending on the relative values of X and Y and the "cost" parameter a, it may or may not be possible to find a profitable value for π at all.

RANDOMIZED COMMITMENTS

Having found a rationale for a "fractional threat," we can inquire whether the tactic of "unconditional commitment," too, is one that in certain cases can advantageously be made less than certain. As indicated in Chapters 3 and 5,[6] a pure commitment — that is, a definite commitment to a pure strategy — is equivalent to "first move" in a two-person, two-move game in which one would otherwise have to move second; it is a means of obtaining the equivalent of first move. We have to relax that interpretation if we suppose that Row, who has second move in the game but who has the option to commit himself ahead of time, commits himself to a 50-50 chance of choosing row i or ii. To

[6] Pp. 47, 122.

do this one must retain the right to move second, exploiting only the right to commit oneself ahead of time; if one had actually to move first, by a definite choice, the possibility of a randomized commitment would be lost. (The randomized commitment is equivalent to a "first move" determined by a random device with odds set by the player, with the odds but not the actual move known to the other player before his own move.)

The same payoff matrix (Fig. 1) can be used to illustrate this situation if we change the rules of the game to permit Row an *unconditional* commitment prior to Column's choice but not permitting him to make his choice depend on Column's. A firm commitment to ii induces a choice of column I but is wasted because the lower left cell — to which Row is now committed — contains no reward. Row's problem is that he needs row ii to induce Column into I, but he needs row i to profit from I. A compromise can be achieved by a randomized commitment — a commitment to a randomized choice. If Row is committed to flip a coin (50-50 chance) to select i or ii after Column has chosen, Column will choose I as long as X is greater than 1.[7] In that case Row gets an expected value of 0.5. If Row sets π (the probability of his choosing ii) at just above $1/(1+X)$ he gets the largest expected value consistent with Column's choice of I. (If Column's payoff in the lower left cell differs from zero, say 0.5 or -0.5, the formula for optimum value of π differs somewhat.) If Row's payoff in the lower left cell were -1, no commitment with a greater than 50 per cent chance of ii would serve. And if that payoff were $-X$ or worse, no probability mixture of i and ii would work; any mixture with π large enough to induce column I would be too large to yield Row a positive expected value.

There is another rationale for a fractional commitment. In the case just discussed, it was Row's own preference for the upper cell in I that led him to minimize the value of π. In Fig. 18 it is *Column's* motivation that demands some chance of row i, that is, a fractional value of π. In this case, a firm commitment to row ii induces Column to choose II; a firm commitment to i induces

[7] That is, as long as the payoff to Column in the lower right cell falls short of his payoff in the upper left as much as the payoff in the upper right exceeds the upper left. See the earlier footnote on the scaling of payoffs.

	I	II
i	4 2	1 1
ii	0 3	2 2

FIG. 18

Column to choose I; no commitment at all leaves Column preferring II; a threat to choose i unless Column chooses I will be ineffective unless Row promises to abstain from choosing ii. In all of these "pure-strategy" cases, Row ends up with a score of 2. He can, however, do slightly better with a mixed commitment. He can, because he and Column are both attracted to column I, disagreeing only over the choice of Row in that column. If he offers Column a 50-50 chance between rows i and ii, Column gets an expected value of 2 in the first column, of 1.5 in the second, and chooses the first. This leaves Row an expected value of 2.5. Since Row has a preference for ii, he wants the highest probability of that row consistent with the need to provide Column with a preference for column I. That is, he wants the largest value of π for which (in the matrix shown)

$$4(1-\pi) > (1-\pi) + 2\pi$$

or

$$3/5 > \pi.$$

This particular mixed commitment can be called a *combination* of a fractional threat with a fractional promise. Row, in effect, "threatens" a relatively high probability of i in the event that II is chosen and "promises" it if I is chosen.

He could do even better if he could make π *conditional* on Column's choice. Any probability up to 0.75 for row ii, conditional on a choice of column I, is a sufficient inducement if it is certain that Row will retaliate for column II with row i. But if he is limited to making his threat no worse than his promise is good — if he has to attach the same probability to both of them — the upper limit to an effective value of π is 0.6, with an expected value to Row of 2.6 (and of 1.6 for Column). With a separate π for the promise, the upper limit is 0.75 for an expected payoff of 2.75 (and only 1.0 for Column).

8

THE THREAT THAT LEAVES SOMETHING TO CHANCE

It is typical of strategic threats that the punitive action — if the threat fails and has to be carried out — is painful or costly to both sides. The purpose is deterrence *ex ante*, not revenge *ex post*. Making a credible threat involves proving that one would have to carry out the threat, or creating incentives for oneself or incurring penalties that would make one evidently want to. The acknowledged purpose of stationing American troops in Europe as a "trip wire" was to convince the Russians that war in Europe would involve the United States whether the Russians thought the United States wanted to be involved or not — that escape from the commitment was physically impossible.

As a rule, one must threaten that he *will* act, not that he *may* act, if the threat fails. To say that one *may* act is to say that one *may not*, and to say this is to confess that one has kept the power of decision — that one is not committed. To say only that one *may* carry out the threat, not that one certainly will, is to invite the opponent to guess whether one will prefer to punish himself and his opponent or to pass up the occasion. Furthermore, if one says that he may — not that he will — and the opponent fails to heed the threat, and the threatener chooses not to carry it out, he only confirms his opponent's belief that when he has a clear choice to act or to abstain he will choose to abstain (consoling himself that he was not caught bluffing because he never said that he would act for sure).

There are threats of this kind nevertheless that may be effective in spite of this loophole. They can work, however, only through a process that is a degree more complicated than firm

commitment to certain fulfillment. Furthermore, they may arise inadvertently and may entail unintended behavior. For this reason they are less likely to be recognized and understood.

The key to these threats is that, though one may or may not carry them out if the threatened party fails to comply, *the final decision is not altogether under the threatener's control.* The threat is not quite of the form "I may or may not, according as I choose," but, has an element of, "I may or may not, and even I can't be altogether sure."

Where does the uncertain element in the decision come from? It must come from somewhere outside of the threatener's control. Whether we call it "chance," accident, third-party influence, imperfection in the machinery of decision, or just processes that we do not entirely understand, it is an ingredient in the situation that neither we nor the party we threaten can entirely control. An example is the threat of inadvertent war.

THE THREAT OF INADVERTENT WAR

The thought that general war might be initiated inadvertently — through some kind of accident, false alarm, or mechanical failure; through somebody's panic, madness, or mischief; through a misapprehension of enemy intentions or a correct apprehension of the enemy's misapprehension of ours — is not an attractive one. As a general rule one wants to keep such a likelihood to a minimum; and on the particular occasions when tension rises and strategic forces are put on extraordinary alert, when the incentive to react quickly is enhanced by the thought that the other side may strike first, it seems particularly important to safeguard against impetuous decision, errors of judgment, and suspicious or ambiguous modes of behavior. It seems likely that, for both human and mechanical reasons, the probability of inadvertent war rises with a crisis.

But is not this mechanism itself a kind of deterrent threat? Suppose the Russians observe that whenever they undertake aggressive action tension rises and this country gets into a sensitive condition of readiness for quick action. Suppose they believe what they have so frequently claimed — that an enhanced status for our

retaliatory forces and for theirs may increase the danger of an accident or a false alarm, theirs or ours, or of some triggering incident, resulting in war. May they not perceive that the risk of all-out war, then, depends on their own behavior, rising when they aggress and intimidate, falling when they relax their pressure against other countries?

Notice that what rises — as far as *this* particular mechanism is concerned — is not the risk that the United States will *decide* on all-out war, but the risk that war will occur whether intended or not. Even if the Russians did not expect deliberate retaliation for the particular misbehavior they had in mind, they could still be uneasy about the possibility that their action might precipitate general war or initiate some dynamic process that could end only in massive war or massive Soviet withdrawal. They might not be confident that we and they could altogether foretell the consequences of our actions in an emergency, and keep the situation altogether under control.

Here is a threat — if a mechanism like this exists — that we *may* act massively, not that we certainly will. It could be most credible. Its credibility stems from the fact that the possibility of precipitating major war in response to Soviet aggression is not limited to the possibility of our coolly deciding to attack; it therefore extends beyond the areas and the events for which a more deliberate threat is in force. It does not depend on our preferring to launch all-out war, or on our being committed to, in the event the Russians confront us with the *fait accompli* of a moderately aggressive move. The final decision is left to "chance." It is up to the Russians to estimate how successfully they and we can avoid precipitating war under the circumstances.

The threat — if we call this contingent-behavior mechanism a "threat" — has some interesting features. It may exist whether we realize it or not. Even those who have doubted whether our massive-retaliation threat was a potent deterrent to *minor* aggression during the last several years, but are perplexed that the Russians have not engaged in more mischief than they have, can note that the threat we voiced was backed by an additional implicit threat that we might be triggered by Soviet actions in spite of ourselves. Furthermore, even if we prefer not to incur even a

small probability of inadvertent war, and would not use this mechanism deliberately, the "threat" in question may be a by-product of other actions that we have a powerful incentive to take. We may get this threat whether we like it or not when we (and the Russians) take precautions commensurate with a crisis; knowing this, the Russians may have to take the risk into account. Finally, the threat is not discredited even if the Russians accomplish their purpose without triggering war. If the Russians estimate that the chance of inadvertent war during a particular month rises from very small to not-so-small if they create a crisis, and they go ahead anyway, and no major war occurs, they still have little reason to suppose that their original estimate was wrong, and little reason to suppose that repetition would be less risky, any more than a person who survives a single play of Russian roulette should decide it isn't dangerous after all.

LIMITED WAR AS A GENERATOR OF RISK

Limited war as a deterrent to aggression also requires interpretation as an action that enhances the *probability* of a greater war. If we ask how the Western forces in Europe are expected to deter a Russian attack or to resist it if it comes, the answer usually runs in terms of a sequence of *decisions*. In case of attack on a moderate scale, we could make the decision to fight limited war; it would not be a decision to proceed with mutual annihilation. If we can resist the Russians on a small scale, they must either give up the idea or themselves take a step upward on the scale of violence. At some point there is a discontinuous jump from limited war to general war, and we hope to confront *them* with that choice. If this is not the typical sequence of decisions envisaged, it at least seems typical in one respect: it involves *deliberate* decisions — decisions to take an action or to abstain from it, to initiate a war or not to, to step up the level of violence or not to, to respond to a challenge or not to.

But another interpretation can be put on limited war. The danger of all-out war is almost certainly increased by the occurrence of a limited war; it is almost certainly increased by an enlargement of limited war. This being so, the threat to engage in

limited war has two parts. One is the threat to inflict costs directly on the other side, in casualties, expenditures, loss of territory, loss of face, or anything else. The second is the threat to expose the other party, together with one's self, to a heightened risk of general war.[1]

Here again is a threat that all-out war *may* occur, not that it certainly will occur, if the other party engages in certain actions. Again, whether it does or does not occur is not a matter altogether controlled by the threatener. Just how all-out war would occur — just where the fault, initiative, or misunderstanding may occur — is not sure. Whatever it is that makes limited war between great powers a risky thing, the risk is a genuine one that neither side can altogether dispel if it wants to. The final decision, or the critical action that initiates an irreversible process, is not something that should necessarily be expected to be taken altogether deliberately. "Chance" helps to decide whether general war occurs or not, with odds that are a matter of judgment based on the nature of the limited war and the context in which it occurs.

Why would one threaten limited war rather than all-out war to deter an attack? First, to threaten limited war — according to this analysis — is to threaten a risk of general war, not the certainty of it; it is consequently a lesser threat than the massively retaliatory threat and more appropriate to certain contingencies. Second, it has the advantage, in case the enemy misjudges our intentions or commitments, of an intermediate stage: we can *engage* in limited war, creating precisely the risk for both of us that we threatened to create, without thereby making general war the price we both pay for the enemy's mistaken judgment. We pay instead the lesser price of a risk of general war, a risk that the enemy can reduce by withdrawal or settlement.

Third, in case the enemy is irrational or impetuous, or we have misjudged his motives or his commitments, or in case his aggressive action has gotten up too much momentum to stop, or his actions are being carried out by puppets or satellites that are

[1] The same point is stressed by Glenn H. Snyder, "Deterrence by Denial and Punishment" (Research Monograph No. 1: Princeton University Center of International Studies, January 2, 1959), pp. 12, 29.

beyond his immediate power to control, there is some prudence in threatening risk rather than certainty. If we threaten all-out war, thinking it not too late to stop him, and it is, we must either go ahead with it or have our threat discredited. But if we can threaten him with a one-in-twenty chance of all-out war in the event he proceeds, and he does proceed, we can hold our breath and have nineteen-to-one odds of getting off without general war. Of course, if we scale down the risk to us, we scale it down to him too; it may degrade the threat to put too much safety in it. But in cases where there is danger that we completely misjudge the enemy's commitment to an action, or completely misjudge his ability to control his own agents, allies, or commanders, the more moderate risk may deter anything that is still within his control.

If we give this interpretation to limited war, we can give a corresponding interpretation to enlargements, or threats of enlargement, of the war. The threat to introduce new weapons into a limited war is not, according to this argument, to be judged solely according to the immediate military or political advantage, but also according to the deliberate risk of still larger war that it poses. Just as a moderate limited war may increase by a large factor the likelihood of major war within the next thirty days, so a progression from conventional to novel weapons may raise that probability by another factor.

We are led in this way to a new interpretation of the "trip wire." The analogy for our limited-war forces in Europe is not, according to this argument, a trip wire that certainly detonates all-out war if it is in working order and fails altogether if it is not. What we have is a graduated series of trip wires, each attached to a chance mechanism, with the daily *probability* of detonation increasing as the enemy moves from wire to wire. The critical feature of the analogy, it should be emphasized, is that whether or not the trip wire detonates general war is — at least to some extent — outside our control, and the Russians know it.

The same interpretation might be true of Quemoy. One can argue that the Chinese or Russians were deterred by the prospect of major war, not just by the prospect of losing a limited war or winning one at excessive cost. Even if they were convinced that we would exercise every skill and caution to keep a war

limited, and they were prepared to exercise skill and caution themselves, they may simply have felt that the process that leads to bigger and bigger wars is not one that they or we fully understand or can foresee, and that the risk, though numerically small, was appreciable.

RISKY BEHAVIOR IN LIMITED WAR

If one of the functions of limited war, then, is to pose the deliberate risk of all-out war, in order to intimidate the enemy and to make pursuit of his limited objectives intolerably risky to him, the usual precepts for behavior in limited war need revision. The supreme objective may not be to *assure* that it stays limited, but rather to keep the risk of all-out war within moderate limits *above zero*. At least this may be the strategy for the side that is in danger of "losing" a limited war. The less likely it is that the enemy's aggressive advances can be contained by limited and local resistance, the more reason there may be to fall back upon the deliberate creation of mutual risk. (Alternatively, the more the aggressor can design his advances so that even local resistance seems fraught with explosive potential, the less attractive local resistance will seem.)

Deliberately raising the risk of all-out war is thus a tactic that fits the context of limited war. Of course, one cannot raise the risk just by saying so. One cannot just announce to the enemy that yesterday one was only about 2 per cent ready to go to all-out war but today it is 7 per cent and they had better watch out. One has to take actions that — assuming he and his adversary continue to be just as concerned and careful to keep the war limited — leave everyone just a little less sure that the war can be kept under control.

The idea is simply that a limited war can get out of hand by degrees. At any point one has some notion or sensation of how much "out of control" it is. And various actions — innovations, breaches of limits, manifestations of "irresponsibility," challenging and assertive acts, adoption of a menacing strategic posture, adoption of headstrong allies and collaborators, spoofing and harassing tactics, introduction of new weapons, enlargement of

troop commitments or the area of conflict — tend to raise almost anyone's judgment of how much "out of control" the situation is. To share such an increase in risk with an enemy may provide him an overpowering incentive to lay off. Preferably one creates the shared risk by irreversible maneuvers or commitments, so that only the enemy's withdrawal can tranquilize the situation; otherwise it may turn out to be a contest of nerves.

REPRISAL AND HARASSMENT

Limited local war is not the only context in which deliberately risky behavior may be used as a type of threat. Between the threats of massive retaliation and of limited war there is the possibility of less-than-massive retaliation, of graduated reprisal. Few serious analyses of war of limited reprisal have been published.[2] The idea that one might "take out" a Russian city if Soviet troops invade a country, and keep "taking out" one every day until they quit, has been occasionally adverted to journalistically but not systematically explored. Similar in spirit is the idea of hostile action on a small scale — sinking ships, blockading ports, jamming communications, or whatever it may be.

There are a number of Russian actions of an aggressive or hostile sort that might provide neither locale for a limited war nor the dramatic act to trigger massive retaliation: efforts to harass, blackmail, or blockade neutral countries or American allies, a peacetime campaign to jam our early-warning and other radar, tricks with nuclear weapons as part of a war of nerves, instigation of sabotage in NATO countries, flagrant support of insurrection, or even the use of unaccustomed violence in quelling disturbances within their own satellites. It may do little good to combat these actions by like measures of our own; it may also not be wise to insist that we are about to boil over into massive retaliation. If something were to be done, the deliberate creation of a small but appreciable shared risk of general war might be considered. (Or, if not, at least the purpose and significance of

[2] A recent serious discussion is Morton A. Kaplan, "The Strategy of Limited Retaliation" (Policy Memorandum 19 of the Center of International Studies; Princeton, April 9, 1959).

Soviet mischief may need to be interpreted as an effort to intimidate by the creation of a shared risk of general war.)

How do we interpret a dramatic act like, say, limited nuclear reprisal on enemy territory? As in limited war, there again may be two parts to the "cost" imposed on an enemy. One is a direct cost: casualties, destruction, humiliation, or whatever it may be. The other is the created risk of all-out war. Nobody quite knows what happens if one country explodes a nuclear weapon in an enemy country. If the action is recognized as an isolated act, limited in intent, not part of a massive attack nor of a sneak attack against the other's retaliatory capability, the victim may not see wisdom in unleashing all-out war in response to the pain and insult. But, even if he does not, he is likely to do something that in turn will have consequences that may ultimately reach a stage of all-out war. If the response is simply to strike back in like fashion, the process may taper off, or it may explode. So, even if each side prefers to act cautiously, failure to understand completely how each other reacts might bring about a dynamic process that ultimately explodes into all-out war.

The odds may still be against it. Here again we are dealing with an action that *may or may not* bring about general war, the final outcome *not* being under the complete control of the participants, the probability of all-out war being a matter of judgment. To mention these possibilities is not necessarily to propose them, but to indicate how they should be interpreted. The sanction they impose on the victim — one that the threatener shares with him — is the recognizable increase in the likelihood of total war.

RISKY BEHAVIOR AND "COMPELLENT" THREATS

There is typically a difference between a threat intended to make an adversary *do* something (or cease doing something) and a threat intended to keep him from starting something. The distinction is in the timing, in who has to make the first move, in whose initiative is put to the test. To deter by threat an enemy's advance it may be enough to burn the bridges behind me as I face the enemy; to compel by threat an enemy's retreat I have to be committed to move forward, and this requires setting fire to the

grass behind me with the wind blowing toward the enemy. I can block your car in the road by placing my car in your way; my deterrent threat is passive, the decision to collide is up to you. If you, however, find me in your way and threaten to collide unless I move, you enjoy no such advantage; the decision to collide is still yours, and I enjoy deterrence. You have to arrange to *have* to collide unless I move, and that is a degree more complicated.

The threat that compels rather than deters, therefore, often takes the form of administering the punishment *until* the other acts, rather than *if* he acts. This is so because often the only way to become physically committed to an action is to initiate it. Initiating steady pain, even if the threatener shares the pain, may make sense as a threat, especially if the threatener can initiate it irreversibly so that only the other's compliance can relieve the pain they both share. But irreversibly initiating certain disaster, if one shares it, is no good. Irreversibly initiating a moderate *risk* of mutual disaster, however, if the other's compliance is feasible within a short enough period to keep the cumulative risk within tolerable bounds, may be a means of scaling down the threat to where one is willing to set it going. Subjecting the enemy (and oneself) to a 1 per cent risk of enormous disaster for each week that he fails to comply is somewhat similar to subjecting him (and oneself) to a steady weekly damage rate equivalent to 1 per cent of disaster. (The words "somewhat" and "equivalent" may be interpreted very flexibly here.) [3]

"Rocking the boat" is a good example. If I say, "Row, or I'll tip the boat over and drown us both," you'll say you don't believe me. But if I rock the boat so that it *may* tip over, you'll be more impressed. If I can't administer pain short of death for the two of us, a "little bit" of death, in the form of a small probability that the boat will tip over, is a near equivalent. But, to make it work, I must really put the boat in jeopardy; just saying that I may turn us both over is unconvincing.

[3] To initiate risky action, if one cannot initiate it irreversibly, does not necessarily "win" over an opponent: the latter may still hope, by acting firm, to induce the initiator to back down. One still has to win the "war of nerves" if the adversary chooses to play it out for a while. But at least this symmetrical situation replaces one in which the asymmetry favored the opponent, who won by default if neither side acted.

Ideally, for this purpose, I should have a little black box that contains a roulette wheel and a device that will detonate in a way that unquestionably provokes total war. I then set this little box down, tell the Russians that I have set it going so that once a day the roulette wheel will spin with a given probability (numerically specified and known to the Russians) that, on any day, the little box will provoke total war. I tell them — *demonstrate to them* — that the little box will keep running until my demands have been complied with and that *there is nothing I can do to stop it.* Note that I do not insist that I shall *decide* on total war, or initiate it deliberately, if the box hits the critical combination. I leave it all up to the box which *automatically* engulfs us both in war if the right (wrong) combination comes up on any day.[4]

Given the fact that, even if the enemy complies, there is some risk that the box detonates war before he has a chance to collect himself and do our bidding, there is an advantage in making it less than certain that the box will explode on any given day. In ordinary deterrence — where nothing happens *unless* the enemy acts contrary to our demand — to threaten too much may be superfluous but not self-defeating; in the present case — where the threat starts fulfilling itself at a specified rate over time as soon as we commit ourselves to it — too big a threat can defeat its purpose. In this situation the small-probability threat is not just a possible substitute for the large certain threat; it is a superior and necessary alternative.

Take an example. A European country, having acquired a modest nuclear retaliatory force, tells the Russians to get out of Hungary or it will work terrible damage on the USSR. The Russians ignore the threat, since there is no persuasive way for the threatening country to make itself *have to* do anything so suicidal. Alternatively, the country threatens to send a missile a day over the USSR, with a nuclear weapon and a random device that explodes it somewhere over Russia if it hasn't been shot down. The Russians say they do not believe the country would do

[4] The tactic may be the less risky, the more automatic the mechanism is; the more automatic it is, the less incentive the enemy has to test my intentions in a war of nerves, prolonging the period of risk.

it; the country does it. The Russians protest and threaten, a day passes, the country does it again. Maybe one weapon gets through and detonates, maybe several do, maybe none do; if some do, maybe they burst over cities, maybe over populated countryside, maybe over deserted areas. The country keeps it up.

What is the country doing? The principal thing the country is doing — in addition to damaging or humiliating Russia — is incurring a painful risk that both it and Russia (and the rest of the world) will be engaged in all-out war in the near future, a war that neither it nor Russia wants. The country is saying in effect, "If you do not get out of Hungary, we *may* cause an all-out war to occur." By when must the Russians get out? The sooner they get out, the sooner the risk of war (from this cause) will be terminated or reduced. The country applying the pressure is not saying, "Get out or we shall deliberately start a war." The decision is not up to them, and does not depend on their displaying the manifest resolution for a final act. The Russians may suppose that the country concerned will do everything it can to prevent total war; but they also have to recognize that with these things flying around, exploding now and then, and with themselves responding in whatever way they feel obliged to, it is not altogether clear that the country concerned, and the Russians, know how to keep total war from occurring.

This illustration is intended just as an analogy for other actions in which posing a risk of all-out war may not be so recognizable as an integral part of what is happening. To take a more immediate situation, suppose an armored column were sent to Berlin in the event that ground access were denied, or suppose, once a transport squeeze on Berlin became intolerable, troops were sent in to claim and hold a corridor; suppose actions were taken that, whether intended to or not, generated some likelihood of an East German uprising. How do we analyze the nature of the pressure on the Russians? I think the answer is in large part that they are confronted with a risk of a war that both sides badly want not to occur, but that both sides may not be able to prevent. A rationale for direct action, even on a scale that by itself might accomplish little, could be the deliberate creation of a risk that we share with

the Russians, providing them with the option either to terminate the risk by acting or to withdraw to meet our objectives.

This is not the only interpretation of such action, of course. It may be that we could win militarily if the fight stays on a small scale, and that for the Russians to enlarge it would require a discontinuous jump that they would be deterred from taking for fear of provoking a discontinuous response. In that case the initial limited war would contain a "deterrent" threat against enlargement of the war. Even so, an important reason why the threat of even small-scale war might be effective is that such a war promises a small but appreciable increase in the probability of an enormous war, the probability being small enough that the Russians believe the West could bring itself to create it, large enough to make it unprofitable for them to let it occur.[5]

It is worth noting that this interpretation suggests that the threat of limited war may be potent even when there is little expectation that we would win it. In these terms, a limited local war is not just local military action; it contains an element of "retaliation" on the Soviet homeland — not a small *bit* of retaliation, but a small *probability* of a massive war.

BRINKMANSHIP

The argument of this paper leads to a definition of brinkmanship and a concept of the "brink of war." The brink is not, in this view, the sharp edge of a cliff where one can stand firmly, look down, and decide whether or not to plunge. The brink is a curved slope that one can stand on with some risk of slipping, the slope gets steeper and the risk of slipping greater as one moves toward the chasm. But the slope and the risk of slipping are rather irregular; neither the person standing there nor onlookers can be quite sure just how great the risk is, or how much it increases when one takes a few more steps downward. One does not, in

[5] In the author's opinion the dispatch of United States troops to Lebanon in 1958 was not only both risky and successful but successful precisely because of the risk — a risk that the Communists could lessen or aggravate according to their response.

brinkmanship, frighten the adversary who is roped to him by getting so close to the edge that if one *decides* to jump one can do so before anyone can stop him. Brinkmanship involves getting onto the slope where one may fall in spite of his own best efforts to save himself, dragging his adversary with him.[6]

Brinkmanship is thus the deliberate creation of a recognizable risk of war, a risk that one does not completely control. It is the tactic of deliberately letting the situation get somewhat out of hand, just because its being out of hand may be intolerable to the other party and force his accommodation. It means harassing and intimidating an adversary by exposing him to a shared risk, or deterring him by showing that if he makes a contrary move he may disturb us so that we slip over the brink whether we want to or not, carrying him with us.

The idea that we should "keep the enemy guessing" about our response, particularly about *whether* we shall respond, needs an interpretation along these lines. It is sometimes argued that we need not threaten the enemy with the certainty of retaliation or the certainty of resistance, but just scare him with the possibility that we may strike back. This idea may be misconceived if it means confronting the Russians with a possible response that remains for us to decide on, one way or the other. The Russians may guess that after the event we should prefer not to strike back, particularly if they perform their aggression in moderate bites; and if we are unwilling to arrange so that we *have* to strike back, and are even unwilling to *say* that we certainly shall, we may seem to confirm their understanding of what our preference would be if we left ourselves any escape. So, if we are afraid that an absolute commitment to the threat might fail in its purpose and commit us to an action we prefer not to be committed to, there may be little to salvage by trying to persuade the enemy that we just might decide to do it anyway.

But the situation is different if we get into a position where it is clear to the Russians that we are sufficiently involved that, while we probably have a way out, we *may* not. To say that we may or may not retaliate for an invasion of some neutral country, depending on how it suits as at the time, and that we shall not

[6] Children understand this perfectly.

let the enemy make this decision for us, nor let him know just what to expect, may confront the enemy with what appears to be a bluff. But to get so involved in or near a neutral country with troops or other commitments that we are not altogether sure ourselves about whether we could evade a fight in case of invasion, may genuinely keep the enemy guessing.

In sum, it may make sense to try to keep the enemy guessing as long as we are not trying to keep him guessing about our own motivation. If the outcome is partly determined by events and processes that are manifestly somewhat beyond our comprehension and control, we create *genuine* risk for him.

THE IMPERFECT PROCESS OF DECISION

Underlying this threat that one "may" retaliate or precipitate war — the decision being somewhat beyond his control — is the notion that some of the most momentous decisions of government are taken by a process that is not entirely predictable, not fully "under control," not altogether deliberate. It implies that a nation can get even into a major war somewhat inadvertently, by a decision process that might be called "imperfect" in the sense that the response to particular contingencies cannot exactly be foretold by any advance calculations, that the response to a particular contingency may depend on certain random or haphazard processes, or that there will be faulty information, faulty communication, misunderstanding, misuse of authority, panic, or human or mechanical failure.

This idea does not reflect an unusually cynical view of the decision process. In the first place, decisions do have to be taken on the basis of incomplete evidence and ambiguous warning; and it is unreasonable to deny *in principle* the possibility of an irrevocable action taken on a false alarm. (Furthermore, one need not be obsessed with the likelihood of false alarm to recognize that there may be levels below which this particular danger cannot be pushed without incurring other dangers that outweigh it!)

Second, war can occur because both sides become committed to irreconcilable positions from which neither is willing to back down, particularly if backing down requires assuming, even mo-

mentarily, a condition of military vulnerability. And it takes no cynic to recognize that two governments may misjudge each other's commitments.

But in the third place, even an orderly government with responsible, comparatively cool-headed leaders is necessarily an imperfect decision system, especially in crises. This is so for a number of reasons, one of which is that in anything but a completely centralized dictatorship a number of persons participate in a decision, and they do not have identical value systems, judgments of enemy intentions, and estimates of military capabilities. A decision taken quickly in crisis may depend on who is present, on whether particular studies have been completed, on the initiative and forcefulness shown by particular leaders and counsellors who are reacting to a quite unprecedented stimulus. Some parts of the decision may be taken on delegated authority, and the person to whom the decision is delegated cannot necessarily reproduce the decision that would have been reached by a president or premier or cabinet in consultation with congressional or parliamentary leaders. There may even be some necessary contradictions in the decision process, such as constitutional issues that cannot be settled in advance but that make it difficult to prepare fully for certain contingencies because the necessity to break law or precedent can be accepted only implicitly, not explicitly prepared for. Finally, the need to keep secrets puts limits on the amount of advance preparation for contingencies that can be carried out.

For this reason there is no such thing as a "firm" plan, intention, or policy of a government to cover every contingency — even all important foreseeable contingencies. How the considerations add up, what interests are brought to bear, and how the collective decision procedure works in future crises is simply not fully determinable in advance.

If on top of this we recognize that there are ordinary human limitations on the intellectual and emotional ability of governmental decision makers during the conduct of dangerous maneuvers on the brink of war, it ought to be clear that there is such a thing as getting into a situation from which it looks as though the nation may successfully extricate itself but in which there is