some appreciable risk that, try as it does within the limits it allows itself, it may not succeed.

One does not expect a government to call attention to its own failings in this regard and to communicate to an enemy that this incomplete mastery of its own actions is an integral part of its strategy. There are also powerful public-relations reasons for not pointing out to an enemy that one is even slightly susceptible to disastrous errors in judgment and false alarms, or that one is a little unsure how to escape from a risky situation. It is understandable, too, that a government engaged in limited war does not state that it has been attracted to this military action by the possible risk of all-out war that it entails. The point is that these things go without saying.

But the basic idea of a threat that leaves something to chance is important even if we do not consciously use it ourselves, even tacitly. In the first place it may be used against us. In the second place, we may misjudge some of the tactics we do use if we fail to recognize the presence of a risk-of-total-war ingredient that may be a significant part of our influence on the enemy even if we have never appreciated it. If — to take an example — this is an important part of the role of limited-war forces in Europe, our analysis of that role may be seriously mistaken if we do not recognize it. The usual idea that a trip wire either does work or does not work, that the Russians either expect it to work or expect it not to work, is mistaking two simple extremes for a more complicated range of probabilities.

PART IV

SURPRISE ATTACK:
A STUDY IN MUTUAL DISTRUST

9

THE RECIPROCAL FEAR OF SURPRISE ATTACK

If I go downstairs to investigate a noise at night, with a gun in my hand, and find myself face to face with a burglar who has a gun in his hand, there is danger of an outcome that neither of us desires. Even if he prefers just to leave quietly, and I wish him to, there is danger that he may *think* I want to shoot, and shoot first. Worse, there is danger that he may think that *I* think *he* wants to shoot. Or he may think that *I* think *he* thinks *I* want to shoot. And so on. "Self-defense" is ambiguous, when one is only trying to preclude being shot in self-defense.

This is the problem of surprise attack. If surprise carries an advantage, it is worth while to avert it by striking first. Fear that the other may be about to strike in the mistaken belief that we are about to strike gives us a motive for striking, and so justifies the other's motive. But, if the gains from even successful surprise are less desired than no war at all, there is no "fundamental" basis for an attack by either side. Nevertheless, it looks as though a modest temptation on each side to sneak in a first blow — a temptation too small by itself to motivate an attack — might become compounded through a process of interacting expectations, with additional motive for attack being produced by successive cycles of "He thinks we think he thinks we think . . . he thinks we think he'll attack; so he thinks we shall; so he will; so we must."

It is interesting that this problem, though it arises most dramatically in situations that would usually be characterized as conflict, like that between the Russians and us or between the burglar and me, is logically equivalent to the problem of two or more partners who lack confidence in each other. If each is under

some temptation to abscond with the joint assets; if each has a little suspicion that the other may be contemplating the same thing; if each realizes that the other may suspect too, and may suspect himself the object of suspicion; we have a pay-off matrix identical with that of a surprise attack problem. If the heat is on some members of the mob, the rest of the mob may be tempted to rub them out to keep them from squealing, and those in danger may be tempted to squeal in self-defense. So the game structure of "preclusive self-defense" is the same as that of "partnership confidence."

The intuitive idea that initial probabilities of surprise attack become larger — may generate a "multiplier" effect — as a result of this compounding of each person's fear of what the other fears, is what I want to analyze in this chapter. More particularly, I want to analyze whether and how this phenomenon can arise through a *rational* calculation of probabilities or a *rational* choice of strategy, by two players who appreciate the nature of their predicament. The intuitive idea itself, even if misconceived, may be a real phenomenon and motivate behavior; people may vaguely think they perceive that the situation is inherently explosive, and respond by exploding. But what I want to explore is whether this phenomenon of "compound expectations" can be represented as a rational process of decision. Can we build an explicit model of this predicament in which two rational players are victims of the logic that governs their expectations of each other? [1]

INFINITE SERIES OF PROBABILITIES

We might begin by trying to set up the problem as follows. A player operates on a set of probabilities, a potentially infinite series of them. First is the estimated probability, P_1, that the other party "really" prefers to attack, that is, that the other will attack even if he does not fear an attack himself. Second is the probability, P_2, that the other player *thinks* that I "really" prefer to attack him, that is, that I will attack him even if I do

[1] Game theorists will recognize this problem as the nonzero-sum counterpart to what, for zero-sum games, has been called a "dueling game." The nonzero-sum version considered here involves the question of whether to shoot, not when to shoot.

not fear an attack on me. Third is the probability, P_3, that *he* thinks *I* think *he* "really" would; fourth is the probability, P_4, that *he* thinks *I* think *he* thinks *I* "really" would. Fifth, sixth, seventh, and so on are built up by lengthening the train of "he thinks" and "I think" with a separate probability attached to each member of the series. The over-all probability that he will attack is then given by:

$$1 - (1 - P_1)(1 - P_2)(1 - P_3) \cdots .$$

The trouble with this formulation is that nothing generates the series. Each probability is an *ad hoc* estimate, reflecting additional data about the specific information structure of the particular situation. We cannot, starting with a few terms in the series as data, project the rest to infinity, or however far it goes, and operate mathematically on the whole series. The number of terms in such a series can be only as much as a player has time to estimate, or the intellectual stamina to keep in mind, since he has to produce each new term of the series by an independent estimating process. It is true, we might set up particular games with information structures that would yield a formula for the series — for example, a series of spins of a roulette wheel determine whether the other player is told my "true" value system, whether I am told whether he has been told, whether he is told whether I have been told what he was told, and so forth — but these would be special games, and might not illuminate much the general situation we are trying to come to grips with. What we need is a formulation of the problem that permits us to work with a limited number of arbitrary parameters, representing perhaps the initial or "objective" terms in a series, in a context that automatically generates the values of any additional probabilities that may be conceived of through the indefinite reiteration of "He thinks I think." We need to formulate the problem in a way that makes each person's expectations a function of the other's.

A "STRICTLY SOLUBLE" NONCOOPERATIVE GAME

As a first try, we can assign to each of the two players a basic parameter representing the likelihood *that he would attack if*

he should not. The values of these parameters are to be fully known and known to be known by both players. What I mean by "should not" is contained in the following two-part behavior hypothesis.

The first part of our behavior hypothesis is that, if the two players both perceive that a joint policy of no-attack is the best of all possible outcomes for both of them, they will recognize this "solution" and elect to abstain. If, for example, the pay-off matrix is as shown in Fig. 19, each will have confidence in their mutual

	I	II
i	0 0	−.5 .5
ii	.5 −.5	1 1

FIG. 19

confidence and will elect the strategy that yields both players the best possible outcome. This seems to be a fairly modest demand on the rationality of the two players.[2] (It is a questionable one, I suppose, mainly if the superiority of joint no-attack over unilateral surprise attack is small, too small to make both players completely confident that they understand each other. And this possibility — that somebody will be tempted to break discipline just to be on the safe side, or for fear that the other may try to be on the safe side — is allowed for in the second part of the behavior hypothesis, immediately following.)

The second part of the hypothesis is that there is some probability, P_r, for player R, and P_c for player C, that the player will in fact attack when he elects (or should elect) a strategy of no-attack, that is, that his decision will contradict the first part of our hypothesis. This is what was meant by the notion that a player might attack even when he "should not." Just what this

[2] In the terminology of Luce and Raiffa, if the noncooperative game has a "solution in the strict sense," that "solution" is here assumed to prevail. *Games and Decisions*, p. 107. Actually the condition is somewhat stronger here, since the solution is jointly preferred by the two players over all alternative outcomes, not just over all other equilibrium points.

parameter represents we shall leave open: it may be taken to be the probability that the player is irrational, or the probability that the pay-off matrix is misconceived and that he "really" prefers unilateral surprise attack, or the probability that somebody will make a mistake and inadvertently send off the attacking force. This parameter, for each player, is "exogenous" in our decision model: it is a datum provided from outside. It is not generated by the interaction of the two players.

These two parameters, P_c and P_r, are assumed to be plainly visible to the two players; there is nothing secret or conjectural about them. This assumption might seem to beg the question we are trying to answer, but it does not. These two exogenous likelihoods of attack do not by themselves indicate what the probability is that the players will in fact attack. They are only one element. The problem is to see whether, given these basic sources of uncertainty, the interaction of the two players' expectations generate additional motive to attack. We have to put at least *some data* into the problem for expectations and conjectures to work on. The only way to hold the arbitrary inputs to a minimal level is to make these two parameters fully visible; otherwise we must state what each guesses about them, what he guesses the other to guess about them, what he guesses the other to guess that he himself guesses about them, and so on. Again we would have the infinite series of *ad hoc* specifications, with the extra difficulty of dealing with probability distributions of probability distributions. The only way to break clean, and to provide a point of departure for calculating what each should fear the other to fear, is to make this one basic uncertainty for each player a matter of record. What we want to see is how an "objective" source of basic uncertainty generates a superstructure of subjective anxieties about each other's anxiety.

We now have a situation that looks as though it would generate the compound self-defense situation that we spoke of. The first player must consider whether the other player's likelihood of attack is serious; he must also consider that the other player is reciprocally worried. Even a player whose own probability of "irrational" attack is known to be zero must consider that the second may attack not only irrationally but also out of fear that

the first, fearing the second's attack, may try to strike first to forestall it. Thus it does seem as though we might get a compounding of motives.

But we do not. We do not get any regular kind of "multiplier" effect out of this. The probabilities of attack by the two sides do not interact to yield a higher probability, except when they yield certainty. That is, the outcome of this game, starting with finite probabilities of "irrational" attack on both sides, is not an enlargement of those probabilities by the fear of surprise attack; it is either joint attack or no attack. That is, it is a pair of *decisions,* not a pair of *probabilities* about behavior.

We work this problem by recomputing the pay-offs in the original matrix, using the two parameters representing the probability of "irrational" attack. The upper left cell in the matrix stays as it was. The lower right cell has its pay-offs recomputed, as a weighted average of the four cells. For, if both players choose the strategy of no-attack, there is a probability equal to $(1 - P_c)(1 - P_r)$ that no attack will occur, a probability equal to $P_r(1 - P_c)$ that R will attack and C will not, a probability equal to $P_c(1 - P_r)$ that C will attack and R will not, and a probability equal to P_cP_r that both will attack. In the same way, the pay-offs in the lower left cell are a weighted average of the pay-offs in the left column; for if C elects to attack, he certainly does attack, while if R *elects* not to, he actually does or does not with probabilities P_r and $(1 - P_r)$ respectively. Thus with probabilities of irrational attack equal to 0.2 for each player, our original matrix would yield a modified matrix like the one in Fig. 20.[3] With proba-

	I	II
i	0 0	−.4 .4
ii	.4 −.4	.64 .64

FIG. 20

[3] In effect we view the players as choosing — in the language of game theory — between one "pure" strategy and one "mixed" strategy the mixture specified by an autonomous parameter. (They could, of course, further mix the pure and mixed strategies, but in the present instance there is no reason to.)

	I	II
i	0 0	−.1 .1
ii	.4 −.4	.46 −.14

FIG. 21

bilities of irrational attack equal to 0.8 for C and 0.2 for R, we get Fig. 21. And with probabilities of 0.8 apiece for irrational attack, we get Fig. 22.

	I	II
i	0 0	−.1 .1
ii	.1 −.1	.04 .04

FIG. 22

The probabilities of irrational attack in the first of our modified matrices, namely the probabilities of 0.2 for each of the players, prove to be innocuous. That is, they are innocuous *with respect to the choice of strategies*. They yield a new pay-off matrix that still has a "strict solution" in the lower right corner. The *value of the game is reduced* for each player, since there is no escaping those two basic probabilities; but the *contemplation* of the probabilities has not led to their aggravation. Each player has fully taken them into account, has seen that there is still a jointly preferred solution at no-attack, and by the original hypothesis has chosen that strategy.

The last of our modified matrices, with a 0.8 probability for each player, is symmetrical and unstable; each player would now rather attack than hope for joint no-attack, and each knows that the other would too. This is a perverse situation, corresponding to the "prisoner's dilemma" familiar in game theory; the only efficient solution would be a binding agreement to elect no-attack (which still leaves them suffering the reduced value

of 0.04), if binding agreements were institutionally possible and if play were forcibly postponed to give the players a chance to reach such an agreement.[4]

The second of the modified matrices is also unstable, though not in a symmetrical way. Player C's likely irrationality requires player R to anticipate it by attacking in self-defense; player C, knowing this, attacks too.[5]

[4] "Prisoner's dilemma" refers, in game theory, to a configuration of payoffs that gives both players dominant incentives — in the absence of an enforceable agreement to the contrary — to choose strategies that together yield both players a less desirable outcome than if both had made opposite choices. The name derives from the problem of two prisoners, separately interrogated, who may confess to a moderate crime in common or accuse each other of a heavy crime, an accuser going free unless himself accused, the accused one or ones receiving heavy sentences. See Luce and Raiffa, pp. 94 ff.

[5] A somewhat different, and rather interesting, case occurs if we put P_r equal to 0.2 and P_e equal to 0.6. The modified matrix (for R only) is then:

0	.2
−.4	.12

R still has a "dominant strategy" of attack; he does better by attacking, no matter what C does. But in this case, as distinct from the case portrayed in Fig. 19, he is worse off than if neither side had elected to attack. It is C's knowledge of R's dominant strategy that causes them both to get zero. C's "irrationality," expressed in P_e, provides R with a motive for attacking in "self-defense"; but an element in that motive — a small "impurity" in the self-defense motive — is R's possibility of achieving surprise and thus of doing better than just meeting an incoming attack. If R were incapable of surprising C, even when he tried, his pay-off in the upper right cell of the original matrix would be zero, not 0.5, and the modified matrix for R would be:

0	0
−.4	.08

This "worsens" both pay-offs for R in the right-hand column, but the upper more than the lower. It therefore eliminates R's motive to attack, and C knows it, so the outcome is at joint no-attack. Not only, then, may it help both players if the more "irrational" member is incapable of attack; it may even help them both if the "victim" is incapable of achieving surprise even in

The limits to the values of our two parameters, P_r and P_c, beyond which they make the situation unstable and provoke joint attack, are — letting h stand for the value obtained by unilateral surprise attack, $-h$ the value obtained by being attacked while not attacking, 0 the value obtained by simultaneous attack, and 1.0 the value of joint no-attack, for each player —

$$P_c < 1 - h_r,$$
$$P_r < 1 - h_c.$$[6]

Figure 23 illustrates what happens to the "value of the game" for each player, and for each strategy, as one of the P's varies from 0 to 1.0. Putting P_r equal to 0.2, and plotting the values of the game against P_c (based on the matrix of Fig. 19), yields values for C and R as diagrammed. At $P_c = 0.5$, the game be-

---

"self-defense." The condition for this special case, in terms of the parameters used in the next paragraph in the text, is

$$1 - h < P_c < 1/(1 + h).$$

This point can be made more general. Suppose the value of "winning" a war, denoted by h, may exceed 1; if it does, and if it is always a winning strategy to attack when the other does not, both players have dominant strategies at "attack." They both gain zero, when they might have had more if they could have abstained. Suppose, now, that the probability of achieving surprise, and thereby winning, is only Q, so that the expected value to be achieved through unilateral attack is only Qh. If Qh is less than 1, we are back to a matrix with a strictly preferred solution at joint no-attack; and, allowing for the probability of "irrational" attack, the game is stable if $P_c < 1 - Q_c h$ and $P_r < 1 - Q_r h$. Suppose that P_c and Q_c meet the first of these conditions: then it is to R's advantage, as well as C's, that the second condition also be met. If P_r is beyond manipulation, R should wish that Q_r, his own capacity for surprising an enemy, should be less than $(1 - P_r)/h$. Only then can he, and C, gain more than zero. If R can, at his own expense, improve his "enemy's" alert system, or if he can blunt his own surprise capacity in a visible way, to hold Q_r below the limit, he should do so. The principle is the same as that of two partners, somewhat distrustful, who keep two separate private padlocks on the partnership vault. If one could not afford a padlock, the other should provide it to him at his own expense; only then can they do business together.

[6] A more general formula, covering the nonsymmetrical case, and using R_{11}, R_{12}, R_{21}, R_{22} to denote the pay-offs to R in row 1 col 1, row 1 col 2, and so on,

$$\frac{P_c}{1 - P_c} < \frac{R_{22} - R_{12}}{R_{11} - R_{21}}$$

The numerator is the "cost" of erroneously attacking; the denominator is the "cost" of erroneously failing to attack. The criterion is the same, it may be noted, as if P and $(1 - P)$ were sure probabilities rather than probabilities of departure from, and adherence to, a "rational" behavior pattern.

comes unstable, and the value of the game goes to 0 for both players.

That this game does not quite correspond to the original notion of "compounded probabilities" is exemplified by the fact that we

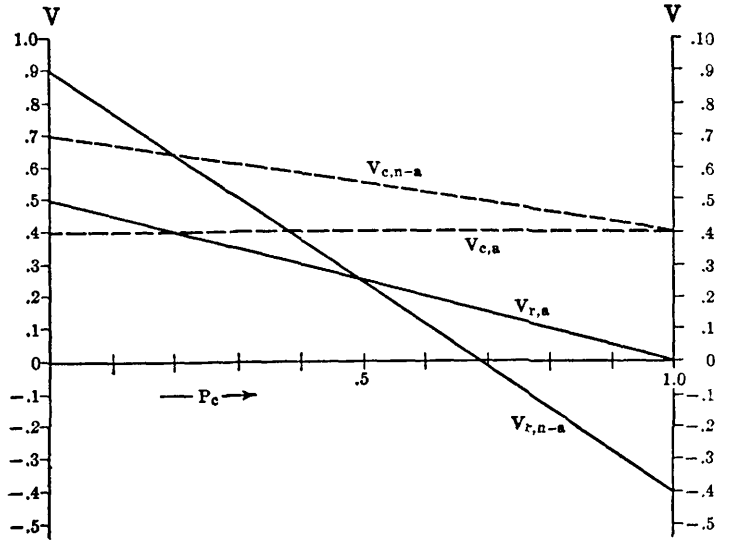

FIG. 23. Value of the game to R and C, as a function of P_c; $P_r = 0.2$. $V_{r,\,n-a}[= 0.9 - 1.3P_c]$: value of game to R, joint strategy of no-attack. $V_{r,\,a}[= 0.5 - 0.5P_c]$: value of game to R, who attacks while C elects not to; $V_{c,\,n-a}[= 0.7 - 0.3P_c]$: value of game to C, joint strategy of no-attack; $V_{c,\,a}[= 0.4]$: value of game to C, who attacks while R elects not to.

can ignore the lesser of the two parameters if they are unequal. If both are below the critical limit, it does not matter what they are; if one is over the limit by ever so little, it makes no difference whether the other is 0 or 1.0. They can thus be potent beyond what they do to the value of joint nonattack, because they can cause the players to shift from a strategy of no-attack to a strategy of attack. But they do so in an all-or-none way. The *likelihood* of attack either is confined to the exogenous likelihoods, or becomes certainty.

THE GAME AS A SEQUENCE OF MOVES IN TURN

We get the same result if we try a game with *moves-in-turn* for the pay-off matrix that we have been using. Suppose R is given a free choice, to attack or not, while C is constrained to wait; and C can attack only *after* R has had an opportunity to make his choice and act on it, and only *if* R has not attacked. We now build further on this game, by letting C have a still earlier choice, preceding R's, so that C gets a turn, then R, then finally C again. We then give R a still earlier turn, so that R chooses, then C chooses, then R chooses, then C chooses (as long as nobody chooses to attack).

What does this game yield? At his last move, C will *elect* not to attack if the matrix is as in Fig. 19; he actually will attack, then, with probability P_c. At his own last move, R knows what C will elect, and makes a predictable choice that depends on P_c. At the preceding move, C knows what R will choose, takes P_r into account and makes a predictable choice. At the move before that, R knows what C will choose to do on both subsequent occasions, takes into account the probability, $1 - (1 - P_c)^2$, that C may attack on either of the next two moves, and makes his own choice in a predictable way. And so on. If each player has n moves, with probability P_r or P_c of irrational attack *at each move,* the outcome depends on whether $P_c = 1 - (1 - P_c)^n$ and $\overline{P}_r = 1 - (1 - P_r)^{n-1}$ meet the conditions derived earlier. If so, each player knows that the other will not subsequently choose to attack, and himself chooses not to at all turns. But if P exceeds the limit for one of them he will prefer to attack and the other knows it, so whoever has first turn attacks at once.

In other words, we are compounding probabilities, but still with an all-or-none effect, and without either player having to *combine* both players' irrationality parameters in the compounding process. Either the probability for at least one of them is big enough and the game long enough to cause the first player to attack, or else no one attacks. And, if we make the over-all probabilities of irrational attack independent of the number of turns, by letting the probability at each turn be equal to

$1 - (1 - P)^{1/n}$, so that the compounded total is just P_c or P_r, the outcome of this game is *independent of the number of turns*. If we think of this game, then, as an *analogy* of the he-thinks-I-think situation, with each turn symbolizing a cycle in the spiral of suspicions, we have a model in which the successive reciprocal fears of what each other fears make no difference: either there is "objective" basis for one of the players to attack, or they abstain.

RECONSIDERATION OF THE PROBLEM

The same seems to be true now if we go back to that burglar downstairs. If he behaves "rationally" as defined in our behavior hypothesis above, he must consider the likelihood that I will shoot him out of sheer preference; and he must consider that I may shoot him if I think there is a strong likelihood that he will shoot me out of sheer preference. But, if we both know what these two basic (exogenous) "likelihoods" are, we need not go any further. Either these basic probabilities are sufficient to make at least one of us shoot to forestall surprise, and hence to make both shoot, so that the second and higher degree fears are superfluous, or else they are insufficient by themselves to make either of us shoot in self-defense, and we know it and have nothing to fear beyond the exogenous likelihoods themselves. If we both can plainly see that neither would be quite induced to shoot solely out of fear of the exogenous probability that the other "really" wants to shoot, then we ought to be able to see that neither needs to fear preclusive action, that neither then needs to fear that the other fears it, and so on.[7]

But a different situation obtains if I shoot not by calculation but by nervousness. Suppose that my nervousness depends on how frightened I am, and my fright depends on how likely I think it that he may shoot me; and suppose he acts the same way. Then when I consider the exogenous probability that he may shoot me out of sheer preference, it makes me nervous; this

[7] For example, if the two could just communicate and check each other's understanding, they could reach an informal agreement not to *elect* to shoot that would leave no incentive to cheat — assuming, still, that the two basic parameters are clearly evident to both of them.

nervousness enhances the likelihood that I may shoot him even though I prefer not to. He sees my nervousness and gets nervous himself; that scares me more, and I am even more likely to shoot. He sees this increment in my nervousness, and matches it with one of his own, scaring me further; and the probability that I will shoot goes up again. Now we can denote each person's nervousness as a function of the other's, and the likelihood of shooting as a function of nervousness, and have a simple pair of simultaneous differential equations that seem to yield precisely the kind of phenomenon we started off to study.[8]

And the reason they do is that this model does not involve *criteria for decision*; that is, it does not involve a behavior hypothesis that tells us which of two strategies a person will select. Instead, our "nervousness model" is one in which people respond to the fear of attack by *a change in the likelihood* that they will themselves attack. Only in this way, by dealing with the *probability* of a player's decision, and not with a *rule* for decision — that is, not with a model in which the player calculates his best strategy and follows it — can we get the kind of "mutual aggravation" phenomenon that I described at the beginning of this chapter.

Now, does this mean that our phenomenon is not one that can be displayed by rational, decisive players? How can we envisage a player reacting to a change in his environment, or to a new bit of information, by *deciding* that he will do something "somewhat more probably" than before? A rational man may be nervous, in which case our theory is physiological rather than intellectual; but can we conceive of the rational game player's taking another look at the burglar and changing the adjustment on his roulette wheel?[9]

[8] There is an important asymmetry in the problem as formulated here. We have allowed for the possibility that one may shoot when he shouldn't and the other knows it — the "nervousness" case — but not for the possibility that one may *not* shoot when he ought to, and the other knows it. (There may be some chance that the burglar has wet ammunition or forgot to load his gun, and I may know that there is such a chance, he may know that I know it, and so forth.) This possibility would apparently be stabilizing, tending to reduce the likelihood of a *decision* to attack as well as the exogenous likelihood of inadvertent or irrational attack.

[9] Note that the usual rationale for a mixed strategy — that is, for rationally

Of course, individual and group decisions may be different in this regard. We could think of a collective decision by vote, with different members having different value systems and hence different thresholds of reaction to the probability of being attacked, so that the size of a vote to attack would be a function of the estimated likelihood of being attacked. If the vote also depends heavily on chance factors, such as absentees on voting day, the *probability* of the required majority in favor of attack becomes a rising function of the probability of the enemy's own decision, which in turn is a function of the first collective player's probability. So we can get the phenomenon we want for "rational" players if we deem rational a collective player that has divergent values and a voting system.

There is, however, a way to adapt our model even to the single, decisive, rational game player. It may be of fairly wide generality in partnership and surprise-attack problems. And it directly involves a significant part of the actual problem of military surprise, namely the dependence of decision on an imperfect warning system, and the possibility of both "type-1" and "type-2" errors in the decision process.

PROBABILITY-BEHAVIOR GENERATED BY AN IMPERFECT WARNING SYSTEM

Presumably the danger of suffering a surprise attack can be reduced by the use of a warning system. But the warning system is not infallible. A warning system may err in either way: it may cause us to identify an attacking plane as a seagull, and do nothing, or it may cause us to identify a seagull as an attacking plane, and provoke our inadvertent attack on the enemy. Both possibilities of error can presumably be reduced by spending more money and ingenuity on the system. But, for a given expenditure, it is generally true of decision criteria that a tightening of the criteria with respect of one kind of error loosens them with respect to the other. To require less evidence of incoming

---

readjusting one's roulette wheel for decision — has no relation to the present case.

attack before "retaliating" is to require more evidence that they are really seagulls for holding back our own planes.

But now we can have a model of a rational decider who responds to an estimate of the probability of being attacked *not* by an overt *decision* to act or abstain, but *by adjusting the likelihood that he may mistakenly attack*. One's response to an increase in the probability of being attacked is to shift the criteria for decision that are used in the warning system in the direction of lesser likelihood of a failure to respond, and hence in the direction of greater likelihood of a false alarm that provokes one's own "retaliation." If each player's response to an increased danger of surprise attack is to enhance his own proclivity toward inadvertent attack, the *probability* of each player's attack is now a rising function of the other's.[10] Such a warning system is the rational, mechanical counterpart of our nervousness in facing the burglar.

To build such a model (symmetrically, for simplicity) we can again let h denote the value of "winning" a war, $-h$ that of "losing" a war, 0 the expected value of simultaneous attack (50-50 chance of winning or losing), and 1.0 the value of no war at all. (This time we can let h exceed 1, as long as $(1 - R)h$ in the matrix below remains below 1. But if "winning" gains a Pyrrhic victory, h will be a small fraction.) We assume that successful surprise wins the war; "successful surprise" means that one attacks when the other does not *and* that the other's warning system fails him. Let R denote the reliability of a player's warning system, that is, the probability that an attack, if it comes, will be identified and surprise forestalled. Then the pay-off matrix is as in Fig. 24.

The probability that a player will attack when he should not, that is, that he will when his rational choice "should" be against attacking (in the sense used earlier), will consist of two parts. One, denoted by A, is the exogeneous likelihood of irrational attack; it excludes the possibility of an attack provoked by false alarm. The probability of an attack through false alarm is de-

[10] As noted below, this is not *necessarily* so; if increased danger of being attacked is associated with reduced vulnerability of the enemy to surprise attack, it is possible for one's response to be in the direction opposite to that described in the text.

noted by B. Thus the two types of error in the warning system are represented by B and $(1 - R)$; and the main feature of the model is that $B = f(R), f'(R) > 0$. That is, the more we reduce $(1 - R)$ as a source of error, the more we increase B, and vice versa.

	I	II
i	0 0	$-(1-R_c)h$ $(1-R_c)h$
ii	$(1-R_r)h$ $-(1-R_r)h$	1 1

FIG. 24

Each player's strategy choice concerns the pair of values for B and R that will minimize his expected losses, that is, maximize the expected value of the game for him. Letting V_r denote the expected value of the game for R, the warning-system problem for R is to choose the pair of values for R and B, consistent with $B = f(R)$, that maximizes [11]

$$\begin{aligned} V_r &= (1 - P_c)(1 - P_r) + P_r(1 - P_c)h(1 - R_c) \\ &\qquad - P_c(1 - P_r)h(1 - R_r) \\ &= (1 - A_c)(1 - B_c)(1 - A_r)(1 - B_r) \\ &\quad + (A_r + B_r - A_rB_r)(1 - A_c)(1 - B_c)h(1 - R_c) \\ &\quad - (A_c + B_c - A_cB_c)(1 - A_r)(1 - B_r)h(1 - R_r). \end{aligned}$$

Additionally, pursuant to the earlier matrix analysis, R should examine the resulting "modified" pay-off matrix that results from using these "optimal" values of R_r and B_r, together with the observed (or expected optimal) values of R_c and B_c, to see whether joint no-attack is still the jointly preferred outcome. The conditions for a joint preference at no-attack, with optimally adjusted warning systems, would be:

[11] It is assumed for convenience of illustration that an inadvertent attack due to false alarm is the same kind of attack as a premeditated attack, with the same likelihood of achieving surprise. Also, we are ignoring the time dimension of B, which probably ought to be thought of as the probability of false alarm per unit of time, while $(1 - R)$ is the probability of error per incoming attack, and A might have some of both elements. Thus the time horizon is assumed fixed in this model.

$$P_c = (A_c + B_c - A_cB_c) < \frac{1 - h(1 - R_c)}{1 - h(R_r - R_c)},$$

$$P_r = (A_r + B_r - A_rB_r) < \frac{1 - h(1 - R_r)}{1 - h(R_c - R_r)}.$$

With symmetry, the denominators in the right-hand terms become just 1.

Actually, as will be seen below, this second examination may be unnecessary; for certain behavior hypotheses, "optimal" adjustment of R and B (for any value short of $R = 1$) requires that the conditions for stability of the modified matrix be met.

It remains to be specified how the players behave. Broadly speaking, we can make either of three hypotheses, corresponding more or less to the difference between "parametric behavior," a "tacit game," and a "bargaining game."

DYNAMIC ADJUSTMENT (PARAMETRIC BEHAVIOR)

First we may try supposing that each player takes the probability of being attacked as given, that is, as a parameter and not a variable in his own loss function, and does the same with the reliability of his opponent's warning system. That is, he directly *observes* the values of his opponent's B and R, and selects the pair of values for own B and R that minimize his expected losses. This assumption tends to make each person's choice of B a rising function of the probability that the other will attack. (It only "tends to," since there is a possibility that the corresponding change in the other's R provides an offsetting inducement, as mentioned below.) If we think of the two players as continually adjusting their values of B and R, each with an eye on the other's B and R, but always responding parametrically to the current probability of being attacked and not projecting the other's behavior as a function of his own, we get a simple dynamic "multiplier" system — stable or explosive depending on the parameter values and shape of the f function. We can express each player's optimum value of B as a function of the other's, solve the two equations, and deduce the stability conditions for the equilibrium. We can also compute "multipliers" relating each

player's changes of B and R to shifts in the f function or to changes in the A parameters.

Explicitly, to find the "parametric-behavior" function for player R we maximize V_r with respect to R_r, subject to $B_r = f(R_r)$ but treating B_c and R_c as fixed. Using the formula given earlier for V_r, we get

$$f' = \frac{P_c h(1 - B_r)}{(1 - P_c)\,[1 - h(1 - R_c)] - P_c h(1 - R_r)}$$

and, for $h(1 - R_c) < 1 > h(1 - R_r)$, $f'' > 0$.

Since f' is presumed positive, the denominator must be positive if V_r is to be maximized with $R < 1$; but the condition that the denominator be positive is precisely the condition that P_c must meet in order that player R still prefer joint no-attack. Thus, if both players have optimal adjustments with $R < 1$, those optimal values of R and B are also perforce consistent with joint preference at no-attack.

The relation of B_r to B_c under this behavior hypothesis, that is, the slope of the resulting function that yields R's optimal B-value for given values of B_c, is obtained by differentiating both sides of the above equation:

$$\frac{dB_r}{dB_c} \text{ (along player R's behavior function)} = \frac{dB_r}{dR_r}\frac{dR_r}{df'}\frac{df'}{dB_c}$$

$$= \frac{f'}{f''}\frac{df'}{dB_c} = \frac{f'}{f''}\left(\frac{\partial f'}{\partial B_c} + \frac{\partial f'}{\partial R_c}\frac{dR_c}{dB_c}\right)$$

$$- \frac{f'}{f''}\left(\frac{\partial f'}{\partial B_c} + \frac{\partial f'/\partial R_c}{\phi'}\right),$$

where $B_c = \phi(R_c)$ denotes the corresponding function for player C.

Since $\partial f'/\partial R_c$ is negative, small values of ϕ' may make player R's dB_r/dB_c negative; it does so by raising the "cost" of inadvertent attack enough to outweigh the increase in the risk of being attacked. In other words, B_r is a function not just of B_c but of $\phi(B_c)$ as well; B_r tends to be increased for a rise in B_c but lowered for a rise in R_c, while B_c and R_c rise together as we consider moving out the B_c axis.

A stable equilibrium requires that player R's dB_r/dB_c and C's dB_c/dB_r should have a product less than 1, that is, that with B_r measured vertically and B_c horizontally, C's curve should intersect R's from below. The general "multiplier" expression relating changes in the B's and R's to shifts in the functions (or to changes in the values of the A's) contains 1 minus this product in the denominator.

As remarked earlier, the denominator in the expression for f' disappears, and R_r, B_r, and f' rise sharply, as h approaches the condition for an unstable matrix. (Actually, stability of the matrix game, as distinct from stability of a parametric-behavior equilibrium, is not a relevant concept for the parametric-behavior hypothesis; to contemplate the matrix and to anticipate the other's action is to project his behavior, not to observe it and adapt to it.)

It may also be noted that player R may ignore A_r in his calculations. It drops out of the formula for optimum B_r and R_r. Intuitively, this is because the only contingency in which *either* the value of R_r or the value of B_r can make any difference is the contingency that R *not* launch "irrational" attack; if he does, B and R are irrelevant to him. (However, A_r does affect the condition for a stable matrix, since it does enter into the condition that P_r must meet. So in projecting C's adjustment, R would have to take A_r into account. But "projecting" C's behavior, rather than just observing B_c and R_c continuously, would make R's behavior nonparametric, contradicting the present hypothesis. If player R were considering the value of spending money to improve his warning system, A_r would affect the calculation since it affects the probability that the system makes any difference; this consideration is outside the present model.)

A TACIT GAME

We can make another behavior hypothesis, which may lead to the same result. Instead of supposing that each player sees how the other's R and B are adjusted, takes them as given, and responds to them; we can suppose that each player knows the technological opportunities of the other player — the functional

relation between R and B for the other player — but cannot reliably observe how the other has adjusted the values of R and B. That is, each understands the mechanics of the other's warning system, but can never be sure just what instructions the other has given on how to interpret the evidence that comes in over the system — the other's decision rule. This hypothesis yields us a noncooperative game, in which each player must choose a value for B (that is, for R), not knowing what value the other has chosen but knowing the other's pay-off matrix.

In this case, we have a pay-off matrix with an "equilibrium point" at precisely the point, if any, where the parametric-behavior hypothesis yielded a stable equilibrium.[12] In other words, what was the "solution" under the parametric-behavior hypothesis is still a candidate for being called a "solution" in the noncooperative form of the game. (In neither case is the equilibrium point necessarily unique. If it is not, the first hypothesis makes the outcome depend on initial conditions and "shocks"; the second tends to complicate the intellectual problem of identifying "solution" strategies.)

This solution, of course, is inefficient for the two players. It is an example of "prisoner's dilemma," mentioned above (p. 214); reciprocal increases in the values of the B's have simply raised the likelihood of attack by each side.[13] There are lesser values for the two B's that would make both parties better off; and if the probabilities of deliberate sneak attack by the two players are equal (A's are equal), an agreement to have no warning system at all, that is, no possibility of false alarm, would be the preferred bargain for the two parties if they were restricted to bargains that gave them identical warning systems.[14]

[12] An *equilibrium point*, in game theory, is a pair of strategies for the two players such that each is optimal vis-à-vis the other. (There may be several such points.)

[13] Economists may find the situation reminiscent of two producers who both allocate their limited productive resources between two commodities. One commodity, "security against false alarm," involves external economies; the other, "security against surprise," involves external diseconomies.

[14] If the A's, B's, and R's are equal, V_r and V_e are equal to $(1 - P)^2$, which has a maximum at $B = 0$. (If B has some minimum value greater than 0, we can attribute it to A.) If B's and R's are equal but the A's are not,

$$dV_r/dB = -2(1 - B)(1 - A_e)(1 - A_r) + (A_e - A_r)(k/f'),$$

A BARGAINING GAME

If we consider the possibility of the two players' negotiating to reduce the sensitivity of their alert systems, in the interest of mutual reductions in B at the cost of smaller R's, and assuming that enforcement of such an agreement were possible, there is no very convincing way of deriving a unique solution without further specification of the bargaining framework. If the *solution* has to be symmetrical and the *game* is symmetrical, that is, if they negotiate over a *common pair of values for R and B*, the result is as just mentioned — o values for B, even if this means o for R, no warning system at all. If warning systems are to be identical, there is some critical difference between the basic probabilities of deliberate sneak attack for the two people (between A_r and A_c) beyond which a side payment would be required for an agreement on the abolition of warning systems.

But, in general, this becomes a wide-open bargaining problem. It is even wider open than the present formulation suggests, since the players may not only manipulate values of R and B but of course can now threaten direct attack, or operate on the institutional arrangements that determine the values of the A's.

There is an enforcement difficulty with any agreement on reduction of values of R and B in the mutual interest; it is that each other's values of R and B may not be observable. They depend — at least to an important extent — on the criteria that will govern future decisions, not solely on the observable, physical mechanics of an alert system. They depend on how long one will wait to be "sure," and on what risks one will accept in an emergency. Furthermore, failure to keep the agreement, if it leads to anything, leads to war itself; so recriminations and damage suits are out of the question if our model represents all-out war

---

which can be positive with A_c greater than A_r and f' small. In this case one of the players — the one with smaller A — has a preference for *some* warning system even if it must be common to both of them, compared with none at all; but it involves lesser values for B and R than parametric behavior (or a non-cooperative game) would lead to, as can be seen by putting the above expression equal to o and comparing the resulting formula for f' with that corresponding to parametric behavior.

rather than a border scrape or a minor transgression of one partner against another.

It might be that $R = B = 0$ is qualitatively observable — the physical "absence" of any system at all. Even this possibility is unavailable, as an enforceable system, if the matrix is unstable with $R = 0$, that is, with $h > 1$. In that case, some "risk" in the form of B is necessary to put the R's safely in the range where $h(1 - R)$ is less than 1.

It may also be difficult to have an agreement that explicitly recognizes the A's, since it may be politically difficult to admit that one's A is above 0.

The players may be driven then to rely on arrangements that either observably blunt their own capacity for surprise or observably improve their own and each other's transformation curves relating R to $(1 - B)$. Both sides may, for example, agree to spend more on the alert system, to make it more efficient; and the richer side may prefer to finance improvements in the other's alert system, rather than leave it in a form that either aggravates the other's sense of insecurity or makes him susceptible to false alarm. An agreement to design forces that have no surprise-attack potential, but instead have improved vulnerability to surprise attack themselves, would seem to be indicated. That is, instead of making R and B the terms of an agreement, they might be forced by the unobservability of R and B to work on the f and ϕ functions themselves, considering each of these functions to involve both one's own alert system and the enemy's (partner's) attack force. (It should be noted, however, that "innovations" in the warning systems — shifts of the f and ϕ functions in the direction of less B for a given level of R and vice versa — are not in all cases stabilizing. Those that raise the *marginal* cost of R may lead to higher values of B; these would be perverse innovations from the point of view of the two players together, analogous to an "improvement" in the prisoners' dilemma matrix that raises each player's pay-off for noncooperative strategies.)

The bargaining-game formulation also lends itself to bargaining-tactic analysis. For example, if one player acts parametrically and the other knows it and takes it into account, the first dis-

plays a "reaction function" [15] which goes into the other's formula for V which the latter tries to maximize. In general, the analysis of "strategic moves" of the kind discussed in Chapters 2, 5 and 7, are relevant to this version of the surprise-attack, partnership-discipline game.

MORE THAN TWO PLAYERS

An interesting variant of the problem would occur if the number of players were increased, or if a third player were brought in as an autonomous agent. To the extent that attack from other quarters must be anticipated, the incentive toward mutual reduction of alert systems is reduced. It remains true, however, that any two players in a larger game can find some advantage in jointly modifying their alert systems, in the direction of lesser danger of false alarm, by taking into account the "external diseconomies" for each other that they leave out of account when behaving parametrically. Two armed watchmen patrolling the same building, each subject to some temptation to shoot on sight, would be better off if they could find some way of reaching an enforceable agreement to be a little less ready to shoot on sight, to reduce the likelihood of shooting each other. (Actually, the two-watchmen problem is a representation of our original model, if we let our original parameters, P_c and P_r, represent the relative likelihoods that a man met in darkness is a burglar rather than the other watchman. We have to introduce some uncertainty about a burglar's behavior — that is, to let him join the game as a rational third participant trying to anticipate the others' decisions — in order to add complications to what we already had.) [16]

[15] Compare the note on p. 151 regarding the concept of "reaction function."

[16] Arthur Lee Burns, of the Australian National University, has discussed some interesting problems of a three-or-more person world. The deliberate provocation of war between two parties, by a mischievous third party, is a possibility when an overt act of ambiguous authorship can be introduced into the reciprocal-suspicion model; and the analysis takes on additional richness when one considers warning systems that, for technical reasons or by reason of joint custody, permit one or both of the central players to witness what is coming in on the other's radar screen. See his "Rationale of Catalytic War" (Center for International Studies, Research Memorandum No. 3; Princeton University, 1959).

10

SURPRISE ATTACK AND DISARMAMENT

"Disarmament" has covered a variety of schemes, some ingenious and some sentimental, for cooperation among potential enemies to reduce the likelihood of war or to reduce its scope and violence. Most proposals have taken as a premise that a reduction in the quantity and potency of weapons, particularly of "offensive" weapons and of weapons that either deliberately or incidentally cause great civilian agony and destruction, promotes this purpose. Some schemes have been comprehensive; others have sought to identify particular areas where the common interest is conspicuous, where the need for trust is minimal, and where a significant start might be made which, if successful, would be a first step toward more comprehensive disarmament. Among these less comprehensive schemes, measures to safeguard against surprise attack have, since the President's first "open-skies" proposal in 1955, come increasingly into prominence.

The focus on surprise attack has not reflected an abandonment of interest in a more ambitious dismantlement of arms; rather it represents the philosophy of picking an area where success is most likely, in order to establish some tradition of successful cooperation. The search for safeguards against surprise attack has generally been considered, in our government and elsewhere, not as an *alternative* to disarmament, but as a *type* of disarmament and a possible step toward more.

Nevertheless, though schemes to avert surprise attack may be in the tradition of disarmament, they represent something of an innovation. The original open-skies proposal was unorthodox in its basic idea that arms themselves are not provocative so long as they are clearly held in reserve — so long as their stance is deterrent rather than aggressive. The proposal was also unortho-

dox in its dramatic reminder that, important as it may be to keep secrets from an enemy and in some matters to keep him guessing about what our plans are, it can be even more important to see that the enemy is *not* left to speculate about our intentions toward surprise attack against him *if in fact we are not planning any such attack*. We are interested not only in assuring ourselves with our own eyes that he is not preparing an attack against us; we are interested as well in assuring *him* through *his* own eyes that *we* are preparing no deliberate attack against *him*.

The importance of not keeping that particular secret has an analogue in our alleged political inability to attack first. As General Leslie R. Groves remarked in a speech, "If Russia knows we won't attack first, the Kremlin will be very much less apt to attack us. . . . Our reluctance to strike first is a military disadvantage to us; but it is also, paradoxically, a factor in preventing a world conflict today." [1] We live in an era in which a potent incentive on either side — perhaps the main incentive — to initiate total war with a surprise attack is the fear of being a poor second for not going first. "Self-defense" becomes peculiarly compounded if we have to worry about his striking us to keep us from striking him to keep him from striking us. . . . The surprise-attack problem, when viewed as a problem of reciprocal suspicion and aggravated "self-defense," suggests that there are not only secrets we prefer not to keep, but military capabilities we might prefer not to have.

Of course, it is even better if the other side does not have them either. So there may be advantages in thinking of the surprise-attack problem as one suitable for negotiation.

The innovation in the surprise-attack approach goes further. It has to do with what the scheme is designed to protect and what armaments it takes for granted. An anti-surprise-attack scheme has as its purpose not just to make *attack* more difficult but to reduce or to eliminate the advantage of striking *first*. It must assume that if the advantage of striking first can be eliminated or severely reduced, the incentive to strike *at all* will be reduced.

[1] *The New York Times*, December 29, 1957, p. 20.

It is widely accepted that the United States has the military *power* virtually to obliterate the USSR, and vice versa. And it is widely accepted that, if either side struck the other a major nuclear blow, the nation so hit would have a powerful *incentive* to strike back with equal or greater force. But, if either side can obliterate the other, what does it matter who strikes first? The answer, of course, is that we are not particularly concerned with outliving the Russians by a day; we are worried about whether a surprise attack might have such prospects of *destroying the power to retaliate* as to be undeterred itself by the threat of retaliation. It is not our existing capacity to destroy Russia that deters a Russian attack against us, but our capacity to retaliate after being attacked ourselves. We must assume that a Russian first-strike, if it came, would be aimed at the very power that we rely upon for retaliation.

There is a difference between a balance of terror in which *either* side can obliterate the other and one in which *both* sides can do it no matter who strikes first. It is not the "balance" — the sheer equality or symmetry in the situation — that constitutes mutual deterrence; it is the *stability* of the balance. The balance is stable only when neither, in striking first, can destroy the other's ability to strike back.

The difference between a stable and an unstable balance is illustrated by another offensive weapon against which no good defense was ever devised.[2] The "equalizer" of the Old West made it possible for *either* man to kill the other; it did not assure that *both* would be killed. The tense consequences of this weapon system can be seen on TV almost any night. The advantage of shooting first aggravates any incentive to shoot. As the survivor might put it, "He was about to kill me in self-defense, so I had to kill him in self-defense." Or, "He, thinking I was about to kill him in self-defense, was about to kill me in self-defense, so I had to kill him in self-defense." But if both were assured of living long

[2] A military historian, commenting on the alleged "historical truth" that there has never yet been a weapon against which man has been unable to devise a counterweapon or a defense, reminds us that "after five centuries of the use of hand arms with fire-propelled missiles . . . no adequate answer has yet been found for the bullet" (Bernard Brodie, *The Absolute Weapon* [New York, 1946], pp. 30–31).

enough to shoot back with unimpaired aim, there would be no advantage in jumping the gun and little reason to fear that the other would try it.

The special significance of surprise attack thus lies in the possible vulnerability of retaliatory forces. If these forces were themselves invulnerable — if each side were confident that its own forces could survive an attack, but also that it could not destroy the other's power to strike back — there would be no powerful temptation to strike first. And there would be less need to react quickly to what might prove to be a false alarm.

Thus schemes to avert surprise attack have as their most immediate objective the safety of weapons rather than the safety of people. Surprise-attack schemes, in contrast to other types of disarmament proposals, are based on *deterrence* as the fundamental protection against attack. They seek to perfect and to stabilize mutual deterrence — to enhance the integrity of particular weapon systems. And it is precisely the weapons most destructive of people that an anti-surprise-attack scheme seeks to preserve — the weapons of retaliation, the weapons whose mission is to punish rather than to fight, to hurt the enemy afterwards, not to disarm him beforehand. A weapon that can hurt only *people*, and cannot possibly damage the other side's striking force, is profoundly defensive: it provides its possessor no incentive to strike first. It is the weapon that is designed or deployed to destroy "military" targets — to seek out the enemy's missiles and bombers — that *can* exploit the advantage of striking first and consequently provide a temptation to do so.

In identifying the surprise-attack problem as the possible vulnerability of each side's retaliatory forces to surprise, we are at the point where measures against surprise attack differ drastically from more conventional notions of disarmament. We are also at the source of a number of anomalies and paradoxes that have to be faced if we are to recognize the virtues and defects of particular schemes and to comprehend the motives behind them. It is at this point, also, that we begin to question whether schemes against surprise attack can be viewed as "first steps" toward more comprehensive disarmament in the traditional sense, or instead are incompatible with other forms of disarmament. Can

measures to protect SAC be viewed as first steps toward its dismantlement? Can we initially take cooperative measures to perfect and safeguard each side's capacity to retaliate massively, in the interest of mutual deterrence, and do it as a step toward eliminating the threat of massive retaliation from a tense and troubled world?

Or should we instead recognize measures to safeguard against surprise attack as a compromise — an implicit acceptance of "mutual deterrence" as the best source of military stability we are likely to find — and a recognition that though we may not be able to replace the balance of terror with anything better, there may be much that we can do to make that balance stable rather than unstable.[8]

Once we have identified the surprise-attack problem as the possible vulnerability of either side's retaliatory force to a first strike by the other, it becomes necessary to evaluate military strength, defensive measures, and proposals for the inspection or limitation of armament, with precisely this type of strategic vulnerability in mind. We do not, for example, assess American and Soviet strategic forces by counting up the bombers, missiles, submarines, and aircraft carriers on both sides, as though we wanted to see who could put on the most impressive peace-time parade. "Who is ahead" in the arms race will usually be: *whoever strikes first*. And if we have to plan on the conservative assumption that the other side will strike first, 200 bombers safe against attack may be worth as much as 2000 that have only a 10 per cent chance of survival.

An assessment of defensive measures also comes out differently if we put primary reliance on deterrence. Chicago cannot be hidden, buried in a blast-proof cavern, or kept 10 miles off the ground; but concealment, dispersal, hard shelter, and airborne alert are meaningful defenses in preserving the deterrent force. An active air defense of Chicago that has only a 50-50 chance

[8] In case the reader feels that the argument presented here is correct in principle but uninteresting in fact because the continuous invulnerability of our retaliatory forces is assured beyond any worry, I should like to refer him to Albert Wohlstetter's cogent discussion in "The Delicate Balance of Terror," *Foreign Affairs*, 37:211–234 (January, 1959).

of saving the city from a multi-megaton bomb would be a discouraging prospect, and we have little promise that we could even do that well; but an active defense that could guarantee the survival of a large fraction of our strategic striking force might be more than enough to guarantee the Russians a prohibitive cost in retaliation. Similarly, a defense of Chicago that requires the enemy to triple the size of his attack may be a poor prospect; it may mean only that he invests in a larger initial attack. But a defense of our retaliatory force that requires the enemy to triple the size of his attack may substantially increase the enemy's difficulty of sneaking past our warning system, and appreciably change his likelihood of successfully precluding retaliation.

The same kind of calculation is pertinent to an evaluation of arms limitations. If we look only at the problem of a Russian attack on American cities, it may seem immaterial to the enemy whether he shoots his ICBM's from close up or from afar; accuracy may not make much difference with a multi-megaton bomb fired at metropolitan areas. But if he is trying to destroy a missile or bomber that has been sheltered deep underground with reinforced concrete, accuracy is no longer superfluous. An average aiming error of two or three miles may be nothing in shooting at a large metropolitan area; an attempt to knock out a hard-sheltered retaliatory weapon may require several missiles to get a direct enough hit. Thus *zonal limitations* on the placement of ICBM's might seem an ineffectual form of disarmament in the conventional sense; but in stabilizing deterrence — in reducing the vulnerability of each side's retaliatory forces to the other's forces — the separation of each side's missile sites from the other's, by reducing accuracy, might make a real difference. (For unsheltered planes or missiles, of course, the city-target analogy is unfortunately pertinent.)

On some questions, emphasis on the surprise-attack problem may lead to a downright reversal of the answer that one would get from more traditional "disarmament" considerations. Consider the case of a limitation on the number of missiles that might be allowed to both sides (if we ever reached the point in negotiations with Russia where an agreement limiting the number of missiles were pertinent and inspection seemed feasible). Suppose we had

decided, from a consideration of population targets and enemy incentives, that we would need a minimum expectation of 100 missiles left over after his first counter-missile strike in order to carry out an adequately punitive retaliatory strike — that is, to deter him from striking in the first place. For illustration suppose his accuracies and reliabilities are such that one of his missiles has a 50-50 chance of knocking out one of ours. Then, if we have 200, he needs to knock out just over half; at 50 per cent reliability he needs to fire just over 200 to cut our residual supply to less than 100. If we had 400, he would need to knock out three-quarters of ours; at a 50 per cent discount rate for misses and failures he would need to fire more than twice 400, that is, more than 800. If we had 800, he would have to knock out seven-eighths of ours, and to do it with 50 per cent reliability he would need over three times that number, or more than 2400. And so on. The larger the initial number on the "defending" side, the larger the *multiple* required by the attacker in order to reduce the victim's residual supply to below some "safe" number.[4]

From this point of view, a limitation on the number of missiles would appear to be more stabilizing, *the larger the number permitted.* This would be so for two reasons. First, the larger the number on both sides, the greater is the absolute number of missiles expected to be left over for retaliation in the event that either side should strike first, and therefore the greater is the deterrence to an attempted first strike. Second, the larger the number of missiles on both sides, the greater must be the absolute and proportionate increase in missiles that either side would have to achieve in order to be capable of assuring, with any specified probability, that the other's left-over missiles would be less than some specified number after being attacked. Thus the difficulty of one side's cheating, by disguising and concealing extra missiles, or breaking the engagement and racing to achieve a dominant number, is more than proportionately enhanced by any increase in the starting figures on both sides. In fact, if the numbers to begin with are high enough to strain the budgetary

[4] This assumes that he fires his missiles all together or that, if he fires successive salvos, he has no means of reconnaissance that lets him know, on successive salvos, which particular missiles have already destroyed their targets

capacities of the two enemies, and within these budgetary capacities the number of missiles is high, stability might be imposed by the economic limitation on what either side could do relative to what it would have to do to achieve mastery.

Here is a case, then, in which an "arms race" does not necessarily lead to a more and more unstable situation. For anything like equal numbers on both sides, the likelihood of successfully wiping out the other side's missiles becomes less and less as the missiles on both sides increase. And the *tolerance* of the system increases too. For small numbers on both sides, a ratio of 2 or 3 to 1 may provide dominance to the larger side, a chance of striking first and leaving the other side a small absolute number for striking back. But if the initial numbers on both sides are higher, it may take a ratio of 10 to 1 rather than 2 or 3 to 1 to have a good chance of striking with impunity. Neither side needs to panic if it falls behind a little bit, and neither has any great hope that it could draw far enough ahead to have the kind of dominance it would need.

This greatly simplified view of a "missile duel" is much too specialized to be a strong argument for arms races rather than disarmament. But it does demonstrate that, within the logic of stable deterrence, and of schemes for the prevention of surprise attack, the question of more vs. fewer weapons has to be analyzed on its merits in individual cases. It is *not* a foregone conclusion that disarmament, in the literal sense, leads to stability.

Our attitude toward missile submarines, and toward the problem of devising submarine-detection techniques, should be much affected by whether we are worried about enemy attack or enemy *surprise* attack. If the submarine proves to be for many years a fairly invulnerable site for anti-population missiles, we should perhaps view it not as an especially terrifying development but as a reassuring one. If in fact the best we can hope for is mutual deterrence and we only want the balance to be stable, then the Polaris-type missile carried by a submarine of great mobility and endurance may be the kind of weapon system that we should like to see in adequate numbers on both sides. If it should prove to be both undetectable and highly reliable, it would have the advantage of not needing to strike first in order to strike at all,

of not fearing that an aggressor might hope to knock out the very forces that were supposed to deter him. True, it might seem more reassuring if we had the power to destroy the enemy's missile subs while he did not have the power to destroy ours; but if the power already exists on both sides, and we cannot wish it away, then the most we can hope for is that this capacity to destroy each other be itself sufficiently indestructible that each side is in fact deterred. From that point of view, we perhaps should not even wish that we alone could have the "invulnerable" nuclear-weapon submarine; if in fact we have either no intention or no political capacity for a first strike, it would usually be helpful if the enemy were confidently assured of this. His own manifest invulnerability to our first strike could be to our advantage if it relieved him of a principal concern that might motivate him to try striking first. If *he* has to worry about the exposure of *his* strategic force to a surprise attack by *us*, *we* have to worry about it too.

These thoughts also affect our attitude toward the search for submarine detection. The Navy is urgently seeking a better system of defense against submarines, and there is no question but that we have to devote ourselves intently to the problem. Yet perhaps we ought simultaneously to *hope* that the problem is insoluble. If it were insoluble (in the relative sense in which a technical problem can ever be insoluble) and submarines were destined to be comparatively safe vehicles for a decade or so, stable deterrence might be technologically possible. If submarines prove to be vulnerable themselves, arms technology is less stable than we hope. We have to *try* to detect submarines, because we cannot afford to let the Russians find a technique that we do not know, and because we have to learn all we can about detection to make our submarines less detectable; but like a person who has entered into an agreement with a partner that he cannot trust, we may search like the devil for a loophole, knowing that our partner is searching just as hard, while hoping that no loophole is to found.[5]

[5] This paper being about principles, not about submarines, I can perhaps be excused for pretending here that undetectability on short notice in the open sea is equivalent to invulnerability.

Once we have pressed the argument this far, we may as well carry it all the way. If our problem is to guarantee to an enemy that we have the ability to strike a punitive blow after being struck ourselves — and to assure him that we know that he knows it so that we are under no temptation to doubt the potency of our own deterrence and strike first — we should find virtue in technological discoveries that enhance the anti-population potency of our retaliatory weapons. If it is logical to take measures to guarantee that a larger proportion of our retaliatory forces could survive a first strike on them, the same logic should make us welcome an increase in the potency of those that do survive. As Bernard Brodie has said, "When we consider the special requirements of deterrence, with its emphasis on the punitive aspect of retaliation, we may find a need even for super-dirty bombs. Since the emphasis must be on making certain that the enemy will fear even the smallest number of bombs that might be sent in retaliation, one wants these bombs to be, and thus to appear before the event, as horrendous as possible." [6]

The novelty of this reasoning disappears as soon as we recognize that the "balance of terror," if it is stable, is simply a massive and modern version of an ancient institution: the exchange of hostages. In older times, one committed himself to a promise by delivering his hostages physically into the hands of his distrustful "partner"; today's military technology makes it possible to have the lives of a potential enemy's women and children within one's grasp while he keeps those women and children thousands of miles away. As long as each side has the manifest power to destroy a nation and its population in response to an attack by the other, the "balance of terror" amounts to a tacit understanding backed by a total exchange of all conceivable hostages. We may not, of course, want to exchange quite *that* many hostages in support of this particular understanding with this particular enemy. But in a lawless world that provides no recourse to damage suits for breach of this unwritten contract, hostages may be the only device by which mutually distrustful and antagonistic partners can strike a bargain.[7]

[6] Bernard Brodie, *Strategy in the Missile Age* (Princeton, 1959), p. 295.

[7] It should be emphasized that I am discussing only the problem of major

This line of reasoning is not simply an enormous rationalization for an arms race. It does indeed suggest that "disarmament" in the literal sense, aimed indiscriminately at weapons of all kinds — or even selectively aimed at the most horrifying weapons of mass destruction — could produce instability rather than stability, and might have to be *completely* successful in order not to be disastrous. Nevertheless, there is an important area of arms limitations that is not only compatible with the foregoing analysis but is suggested by it.

It suggests making a distinction between the kinds of weapons that are peculiarly suitable to the exploitation of a first strike and weapons that are peculiarly suitable to the retaliatory role. At one extreme is the "pure" strike-back type of weapon: the relatively inaccurate vehicle with a super-dirty bomb that can kill just about everything in the enemy's country *except* a well-protected or well-hidden retaliatory force, and that itself is so well-protected or well-hidden as to be invulnerable to any weapons that the other side might possess. Ideally, this weapon would suffer no disadvantage in waiting to strike second and gain no advantage in striking first. At the opposite extreme is a weapon that is itself so vulnerable that it could not survive to strike second, or a weapon so specialized for finding and destroying the enemy's retaliatory forces before they are launched that it would lose most of its usefulness if it were held until the other side has already started. These "strike-first" weapons not only give their possessor a powerful incentive to strike first, and an incentive to jump the gun in the event of ambiguous warning rather than to wait and make absolutely sure; they are a tacit declaration to the enemy that one expects to strike first. They consequently invite the enemy to strike a little before *that* and to act with haste in the event he thinks that we think it's time to act quickly.

Between the extremes of the "pure" strike-first weapon and the

---

surprise attack here. The implications of the "hostage" concept for, say, civil defense policy depends on its relation to other contingencies as well — e.g., limited war, mischief by a third party, less-than-massive retaliation, etc. One of these interrelations between surprise attack and other military contingencies is touched on in the final pages of this chapter.

"pure" strike-back weapon, there are the weapons that can strike first but do not need to, that can survive and serve the retaliatory purpose but that also might have an important effect on the other side's retaliatory forces if used first. Perhaps most weapons fall in this category if reasonable precautions are taken for their protection. So we cannot make a nice distinction between first-strike and second-strike weapons, extolling the one and disparaging the other in our approach to the surprise-attack problem. If we were to consider eliminating all weapons that had any possible effect against the other side's retaliatory forces, or that enjoyed any advantage in being used first, there might not be enough left with which to promise retaliation.[8] But surprise-attack negotiations might usefully concentrate on the opposite extreme.

The most obvious candidates would be exposed, vulnerable weapons. It might seem anomalous to insist to the Russians that they cover any nakedness of their strategic forces, or for them to suggest that we protect better some of our own. More likely would be suggestions to abandon weapons that were provocatively exposed to the other side. Note how different in spirit this would be from the "ban the bomb" orientation. Whatever the propaganda implications of such a topic, it at least has the merit of viewing deterrence as something to be enhanced, not dismantled.

Second, restrictions on the deployment of forces that affect their counter-force potency rather than their counter-population potency might be sought. They will not be sought, however, until there is candid recognition that surprise-attack schemes are to be deliberately aimed at protecting, not degrading, each side's strike-back capability. The discussion above of the effect of range on missile requirements, whatever its specific merits, suggests that this class of limitations is not an empty one.

Third, there may be some useful exploration of cooperative measures, or mutually accommodated modes of behavior, that reduce the danger of war by misapprehension. Even voluntary exchange of information might help, if we and the Russians can unilaterally pick modes of behavior that, when the truth is

[8] Furthermore, we are taking nothing but the surprise-attack problem into account here.

known, are reassuring. This is presumably the idea behind proposals for inspection of air traffic in the north polar area, and there may be some other types of activity in which there could be mutual benefit from some traffic rules. What is attractive about these measures — as about a candid discussion of the evils of strike-first weapon systems — is that they may make possible some understandings that do not have to be embodied in formal agreements, and may facilitate unilateral accommodations on both sides.

Fourth, there may be arrangements to cope with crises and emergencies that threaten to explode into an unintended war. A later section of this chapter discusses this point at some length.

Fifth, there may be measures that, by making surprise less likely, make a first strike less attractive. This point brings us back to the open-skies type of proposal.

Most public discussion of the surprise-attack problem during the last few years has related to measures that might reduce the likelihood of surprise, rather than measures to limit what weapons could do if surprise were achieved. The open-skies proposal was based on the idea that with sufficient observation of each other's military forces neither side could achieve surprise and, lacking the advantage of surprise, would be deterred.

The technical problem of devising a practical inspection scheme that could yield each side adequate warning of an attack by the other has become much more difficult since the first open-skies proposal was made. With hydrogen weapons reducing the number of aircraft that might be needed in a surprise attack, with missiles promising to reduce the total time available between the initial actions in readying a strike and the explosion of weapons on target, and with mobile systems like missile submarines to keep under surveillance, it looks as though pure inspection unaccompanied by any limits on the behavior of the things to be inspected would be enormously difficult or enormously ineffectual. The idea of examining photographs for strategic indications of force movements and concentrations is simply obsolete. The problem now would seem to be one of intensive surveillance of strategic forces by a vast organization that could

transmit authentic messages reporting suspicious activity within at most a few hours, and eventually within a few minutes, in a way that is not intolerably susceptible of false alarms. There is no practical assurance that this could be done.

This does not mean that inspection schemes against surprise attack have no prospect of success. What it means is that a scheme providing for *nothing but inspection* may have very poor prospects. But if one cannot send observers out to follow all the aircraft, missiles, and submarines wherever they go, one can still consider calling the aircraft, missiles, and submarines to assemble where they are more easily watched. If restrictions on the deployment of forces are used to make the task of inspection more manageable, something may be accomplished. But though there may be promise in the idea of combining inspection and weapon limitations, there are also serious problems.

One is a possible incompatibility between the need for inspection and the need for concealment. When missiles become sufficiently accurate, it may become almost physically impossible to protect one's own retaliatory forces by the sheer provision of cement, or, if not impossible, exceedingly costly. Mobility and concealment may then have to be the source of security for the retaliatory forces; if the enemy can hit anything he can locate, and kill anything he can hit, he has to be made unable to locate it. To the extent that he can have our own retaliatory weapons under continuous surveillance he has continuous information on their location.

In other ways an inspection scheme on the scale required for protection against surprise attack might yield excessive information about the disposition of the other's forces and make them more vulnerable. It is widely known, for example, that there was a time when hurricane winds immobilized an extremely large portion of the B-36's that then comprised our principal retaliatory threat. The implications for surprise attack of such an event are evidently very different, depending on whether the enemy knows only in a general way that this kind of thing can happen to us, or instead has definite information when it occurs and knows exactly whether or not he has clear sailing for a few days. Imagine the state of tension that could occur if *either* side's

strategic-force personnel began to suffer a severe epidemic that threatened to immobilize them temporarily before the eyes of the other side's inspection. Much better — if we and they are occasionally to land in a very unalert position for reasons that are impossible to prevent — that neither of us should be in a position to know too much about the other's occasional disabilities.

Finally, while there may be arrangements that have a high probability of providing warning of the enemy's preparation for an attack, the value of the system depends on what we can do if we do get warning. We can send off our own anticipatory strike, hoping to get in first; but this is an unattractive course if the warning is ambiguous. A false alarm then leads to war. And a true one precludes any last-minute deterrence.

At the other extreme we can just wait and "get ready." And if the things we can do to get ready appreciably reduce the likelihood that his attack will succeed — if they raise the likelihood that we can retaliate severely — we may want to make a quick demonstration to the enemy that we are ready, in the hope that our improved posture will deter his final decision.

The important question is what we do that constitutes getting ready. If the answer is simply, "Be more alert," why weren't we more alert in the first place? Most of the obvious things that one would do if he had warning of an attack are things that one probably would like to do perpetually in view of the ever-present possibility of an attack. And if our Strategic Air Command is continually doing its best to reduce the time it takes to get aircraft ready and off the ground in the face of warning, or to keep the doors tightly shut on sheltered aircraft, or to keep aircraft safe in the air in combat-ready condition, there may not be much more they can do on short notice.

Nevertheless, there are things that a nation can do in the face of imminent attack that it could not do continuously and indefinitely. One can evacuate or go underground, but not forever. One can get his retaliatory forces safely off the ground, where they are no longer targets for enemy bombs; but they cannot stay in the air forever. One can put men on twenty-four-hour duty, but not for many days in a row. One can ground all com-

mercial aircraft to raise the reliability of the warning system, but the economic loss might be exorbitant if commercial and private flying were foresworn for all time in the interest of making enemy aircraft more recognizable. There are, in other words, things that one can do to "get ready" in the face of expected attack that one cannot be expected to do continuously.

But there is another question. How long can we keep it up? Suppose we cannot physically keep all aircraft in the air at all times, as is true, and that it may be too costly in all respects (accidents as well as fuel and crews) to keep as many as half of them in the sky on the average, but that a substantial increase in the number aloft can be affected on short notice if a serious warning is received. This might well mean that the enemy would not be deterred by our ordinary posture, but would be deterred by the posture we can adopt when we get warning. Does this mean that he quits as soon as he sees that we are ready? Or might he just wait until the gas is gone, the pilots are tired, and the planes have to come down again? And if so, must we not strike in anticipation?

This problem of "fatigue" is likely to plague any super-alert stance that one can take. The solution is in two parts. First, one must try to design a super-alert response that has good endurance and little fatigue, recognizing that this means compromising its peak effectiveness. Second, and most pertinent to the present subject, one may have to engage in a kind of crash disarmament negotiation with the enemy during the period that one has in fact taken measures to insure his own invulnerability of retaliation. If we can keep up a super-alert for a few days, we have a few days during which to attempt to demand or negotiate some degree of Russian "disarmament" that is both tolerable to them and sufficiently reassuring to us to permit us to return to "normal" rather than to proceed with total war. This might mean devising and instituting a much more ambitious scheme of anti-surprise-attack measures than had been politically feasible during the earlier period. It would mean negotiating not just under the ordinary pressure of knowing that sneak attack is a long-term danger, but doing it with clear notice that if measures to

make successful first-strike impossible have not been devised, agreed upon, and taken by a quick deadline, war by mutual consent has become inevitable.

These reflections do not imply that extra warning would be either useless or embarrassing. What they indicate is that warning by itself may not be enough. Extra warning provides an *opportunity,* but the opportunity has to be exploited with skill. And preparations for what one would do in the contingency may have to be made well ahead of time. There is barely time to deliver an ultimatum to the Russians when we catch them preparing to attack. Deciding what ultimatum would both meet our needs and be tolerable to the Russians is not only intellectually difficult, it is technically difficult, depending on such things as procedures to verify compliance. We could probably deliver an effective ultimatum only if we had planned carefully ahead of time on what it might contain.

There are two quite distinct criteria for judging the efficacy of an inspection system, or for designing the system itself. One is how well the system gets at the truth in spite of efforts to conceal it; the other is how well it helps one to reveal the truth convincingly when it is in his interest to do so. The difference is like that between a scheme for discovering the guilty and a scheme for permitting the innocent to establish innocence. Roughly speaking, one system arrives at a presumption of innocence in a negative way, by an absence of positive evidence to the contrary; the other scheme relies on positive evidence, and is pertinent to the particular situations in which one's own interest is in letting the truth be known.

The difference between these two situations is pertinent to the distinction between a scheme to minimize the fear of *deliberate* surprise attack and a scheme to minimize the fear of inadvertent, or "accidental," or unintended war — the war that results from a false alarm, or from a mistaken evaluation of the other's response to a false alarm, or to a wrong interpretation of a mechanical accident, or to the catalytic mischief of a third party interested in promoting war, or to a situation in which the apprehension by each side that the other may be about to pre-empt explodes by

feedback into a war by mutual panic. In the case of a planned, deliberate, surprise attack, the aggressor has every reason to disguise the truth. But in the case of "inadvertent war," both sides have a strong interest in conveying the truth if the truth can in fact be conveyed in a believable way in time to prevent the other side's mistaken decision.

MISAPPREHENSION OF ATTACK

Consider this question: how would we prove to the Soviet Union that we were not engaged in a surprise attack, when in fact we were not but they thought we might be? How might they prove to us that they were not initiating a surprise attack, if in fact they were not but they knew that we were afraid they might be.

Evidently it is not going to be enough just to tell the truth. There may indeed be some situations in which sheer verbal contact is enough to allay each side's suspicions. If the Russians — just to take a wild example — suffered an accidental nuclear explosion on one of their own bases, it might be helpful to both sides if they could simply reassure us quickly that they knew it was an accident, that they were not interpreting it as a harbinger of an attack by us, and so on. But, in most of the cases that one can imagine, it is insufficient simply to assert that one is not engaging in a strategic strike or that one is not in a menacing posture. There has to be some way of authenticating certain facts, the facts presumably involving the disposition of forces. We would have to prove not only that we were not *intending* to exploit our position, but that our actual position was one that *could not* be exploited to doublecross the enemy if he should take us at our word and restrain his own forces.

MISAPPREHENSION DURING LIMITED WAR

Especially in the course of a limited war one side or the other may take an action that might be misinterpreted as a strategic strike. Suppose, for example, that we used the kinds of aircraft that would alternatively be used in a strike against Russian bases,

and flew them in directions that might be interpreted as aimed at the Soviet Union itself — as might be the case if they were flying from North African bases or the Mediterranean fleet to countries near the southern border of the Soviet Union. Alternatively, suppose that the Soviet aircraft flew a limited war mission that could be interpreted, on the basis of the momentary evidence we might get, as a strike at all of our overseas bases and carriers, but that was actually a limited strike and not part of a general effort to destroy United States retaliatory power.

The question arises whether there are any means by which to reduce the likelihood of misinterpretation in this case, where misinterpretation might lead one side either to take off in anticipated retaliation, to pre-empt as quickly as it could, or to get into a super-alert status that had a high proclivity toward false alarm. One might wish to bend over backwards to demonstrate that complementary actions — actions involving other forces in other parts of the world, that would almost certainly take place if this were an all-out counter-force strike — were in fact not being taken.

RECIPROCAL MISAPPREHENSION

Consider another case that was described by Gromyko at a press conference.

> After all, meteors and electronic interference are reflected on Soviet radar screens, too. If in such cases Soviet aircraft, loaded with atomic and hydrogen bombs, were to proceed in the direction of the United States and its bases in other states, the air fleets of both sides, having noticed each other somewhere over the Arctic region, under such circumstances would draw the natural conclusion that a real attack by the enemy was taking place, and mankind would find itself involved in the whirlpool of atomic war.

Assuming for the moment that a situation like that described might conceivably arise, how might the interacting misapprehensions of both sides be slowed down and reversed? If there were some way of reversing motion on both sides, in a properly phased and authenticated way, a kind of balanced withdrawal by mutual consent might be possible.

The bargaining environment is not a propitious one. At best there would be only hours in which to conduct the negotiations, and at worst no time at all. The requirements for a successful outcome can analytically be divided into two parts. First there has to be discovered some "solution" — some pattern of action that reverses the trend toward mutual attack, and that constitutes a dynamically stable withdrawal to a less menacing alert status, one that yields neither side a dangerous advantage in the process, and that is within the physical capabilities of the forces concerned. The second requirement is that compliance somehow be observable, verifiable, and provable. We cannot carry out our part of the bargain unless we have trustworthy means for monitoring the other side's compliance, and the same is true for them. Conceivably we would have an interest in cheating; but it is overwhelmingly more probable that we should wish in these circumstances for a cheat-proof monitoring system that we could submit to, so that if we did comply with our part of the bargain the other side would have no doubt about it. The problem is essentially one of contract enforcement. And the motivation in this case, for each side, is to convey the truth as best it can if in fact it complies with the plan.

This example not only makes clear the need for some *prior arrangement* for observation and verification, in view of the very short time available for bringing inspectors to the scene; it also demonstrates how important it is to have thought ahead of time about what kind of proposal to make, and to have designed one's own flight plan in a way that could take maximum advantage of any means we might have for deliberately giving the enemy true information in the event it becomes desperately necessary to do so.

This case also may illustrate the difference between the two criteria for reliability of an inspection system. It might be very difficult to design radar that would *always* catch the enemy — and by which he could always catch us — in an attempt at sneak attack; it is quite another question how to design radar so that if we both wished to invite voluntary surveillance we could submit in a convincing way. In one case we are, in effect, evading his radar surveillance as best we can. In the other we may de-

liberately "parade" in front of his radar, or submit to other means of long-distance recognition, as long as he does the same for us.

LONGER-TERM SURVEILLANCE

The difference between these crises and emergency situations and the longer-term problems of policing arms limitations is in the kind of evidence that is required and in the strength of the motivation to provide it. The more "leisurely" process of inspection is generally viewed as depending mainly on *negative evidence,* that is, the *absence of evidence.* One reduces the probability of missing such evidence by enlarging and intensifying the system; and one supposes that the evasion is made difficult by the need to keep activities hidden over a long period. But in a crisis one requires more certain evidence; one does not have time to get leads and follow them up; there is no time to try the system out and enlarge it or intensify it if it does not work. Consequently, a crisis agreement would have to rely on *positive evidence.* Instead of looking for evidence about what the other party is *not* doing, one demands evidence that shows what he *is* doing. And the reason why such evidence might be forthcoming in a crisis is that the motive to provide it — the greater urgency of reaching an understanding or an agreement that depends on it — may be enhanced in such an emergency.

OVERBUILDING THE SYSTEM

For the purpose of being at least somewhat prepared for crises and unforeseen situations, there is a good argument for instituting some flexible stand-by arrangements for communicating with potential enemies and inspecting each other. In particular there is a good argument for overbuilding an inspection system relative to such use as has been agreed on. Having standby capacity to enlarge or intensify the system, or to augment it with additional facilities and inspectors, may have a good deal to do with the usefulness of the system in time of crisis. To put the point differently, we should not judge the reliability and usefulness of a system solely in terms of the motivations of the participants during "normal" operations; we should recognize that occasions may

arise when there is a powerful motive for crash negotiations on arms limitations, at least momentary limitations, with no time available for setting up observation and communication systems *ad hoc*.

To be specific: in the event there should be established an inspection system to monitor an agreement to suspend nuclear tests, we should consider carefully how both sides might take advantage of the inspectors and their facilities in the event of an acute military crisis. The mobility of the inspectors, their location, their communication facilities, their technical training and surveillance equipment, there trustworthiness, and their numbers, should be evaluated and designed not just with nuclear-test detection in mind, but with some view to their serving a desperately critical need for a means of inspection, verification, and communication, in a crisis that threatens both us and the Russians with inadvertent war.

From the foregoing considerations, it is not at all clear that the stability of the balance of terror — the lack of temptation to deliberate surprise attack, and the immunity of the situation to false alarm — will be greatly affected by the military arrangements that we try to work out with the Russians. As nature reveals her scientific and technological secrets over the coming years, we may find that each side (if it does what it ought to do and does it rapidly enough) can substantially assure the invulnerability of its own retaliatory forces irrespective of what the other side does, and assure it in a convincing way, so that a powerfully stable mutual deterrence results. Alternatively, nature may have planted mischievous secrets ahead of us, so that we and the Russians continually find new ways to destroy retaliatory forces at a faster rate than we find new ways to protect them. There is only a hope — no presumption — that even with great ingenuity and the best of diplomacy we and the Russians could find cooperative measures to arrest a trend toward instability. So we may get stability without cooperation, or we may not find it even with cooperation. Still, some kind of cooperation with the Russians, or mutual restraint, formal or informal, tacit or explicit, may prove to make a significant difference in the stabil-

ity of the balance of terror; and the stakes of course are very high. So although we cannot be sure that a deliberate policy of collaborating to make each side's retaliatory forces invulnerable would make any difference, we have to consider that it might and to ask ourselves whether in fact we should want a perfectly stable balance of deterrence if we had the option before us. Would we really be interested in a far-reaching and effective anti-surprise-attack scheme if we knew of one, and if we thought the Russians would accept it?

Although it would be comforting to know that the Russians could not be tempted into a deliberate planned sneak attack, and comforting to know that they were so sure we wouldn't try it that they would never need to jump the gun in panic, it can nevertheless be argued that our ability to deter anything but a major assault on ourselves depends at least somewhat on the Russian belief that we might be goaded into deliberate attack. The Russians might not believe this if their retaliatory forces were substantially invulnerable to a first strike by ours. It can be argued that except under the most extreme provocation we would shrink from any retaliatory strike that had no significant chance of eliminating or softening the Russian return strike. According to this argument, a pair of *invulnerable* SAC's is a pair of *neutralized* SAC's; and while that might be the best kind in a completely bi-polar world, it is a luxury that we could not afford in the existing world — a world in which there is a large "third area" in which we wish to deter Russian aggression by a threat more credible than that of mutual suicide.

Can we threaten to retaliate, not just to resist locally, if the Russians unquestionably possess the military capacity to return us a blow of any size they please? Have the strategic forces any role when each is invulnerable to the other, except to neutralize each other and to guarantee, by their joint existence, their joint disuse?

There is a role. Strategic forces would still be capable of carrying out "retaliation" in the punitive sense. If the threat of knocking out Russian or Chinese cities was originally thought to be potent because of the sheer pain, economic loss, disorganization, and humiliation that would be involved, and not mainly because

the military posture of the enemy in the immediate area of his aggression would be greatly affected, the main ingredient of the threat would still be present even if the other side's SAC were invulnerable.

The threat of *massive* retaliation, if "massive" is interpreted to mean unlimited retaliation, does indeed lose credibility with the loss of our hope that a skillfully conducted all-out strike might succeed in precluding counter-retaliation. But if we were ever to consider limited or graduated reprisals as a means of putting pressure on the Russians to desist from actions intolerable to us, or to consider extending a limited local war inside Russian borders in a way that maintained the pretence of local military action but was really intended to work through the sanction of civilian pain and the threat of more, this kind of retaliatory action, and the threat of it, might enjoy increased credibility with a reduction in the vulnerability of both sides' strategic forces. It does, paradoxically, for the same reason that *all kinds of limited war might become less inhibited as the possibility of all-out surprise attack became unavailable.* The risk involved in a bit of less-than-massive retaliation should be less than it is now because the fear of an *all-out* strike in return should be a good deal less. The fear that our limited retaliation would be mistaken for the first step in the initiation of all-out war should be less; the Russians would have to believe that we were literally prepared for suicide to mistake our limited retaliation for the initial step in mutual obliteration.

This is not to argue that limited retaliation, entailing the risk, if not the certainty, of limited counter-retaliation, cannot lead to total destruction, either slowly or by explosion into greater and greater retaliatory strikes, or would not be frightful to contemplate even if kept limited. The problem of limiting a war of reprisal may be no easier than that of limiting local war, and it may be harder. The argument here, however, does not depend on making an exchange of limited punitive blows appear safe and attractive *compared with limited local war,* but safe enough and attractive enough compared with all-out war to be a credible threat (and not a called bluff) *in any case where we may have to rely on the threat of retaliation.*

The strategic forces would thus be "neutralized" only in respect of potential attacks on each other; they would still possess a punitive role that provides some basis for a deterrent threat. While the threat of *all-out* punishment may lose credibility with the achievement of invulnerability by both sides' retaliatory forces, the threat of limited retaliation may well gain it. Whatever the net effect, we cannot deprecate a world of invulnerable SAC's simply by reference to the need for third-area deterrence; it has to be demonstrated that one particular deterrent threat (the massive one) is more potent than the other (limited) one.

Only an extreme optimist can think that we may ever have a clear choice of accepting or rejecting a scheme that would guarantee to make both sides' retaliatory forces totally and continuously invulnerable. But this question of what would happen to third-area deterrence, and the limited-retaliation possibility that it calls to mind, are pertinent to the question of what we might let ourselves hope for.[9]

[9] For further discussion see T. C. Schelling and Morton H. Halperin, *Strategy and Arms Control*, The Twentieth Century Fund (New York, 1961).

APPENDICES

APPENDIX A

NUCLEAR WEAPONS AND LIMITED WAR

With the development of small-size, small-yield nuclear weapons suitable for local use by ground troops with modest equipment, and with the development of nuclear depth charges and nuclear rockets for air-to-air combat, the technical characteristics of nuclear weapons have ceased to provide much basis, if any, for treating nuclear weapons as peculiarly different from other weapons in the conduct of limited war. It has, of course, been argued that there are political disadvantages in our using nuclear weapons in limited war, particularly in our using them first. Even those who consider a nuclear fireball as moral as napalm for burning a man to death must recognize as a political fact a worldwide revulsion against nuclear weapons.

This Appendix is about another basis for distinguishing between nuclear and other weapons. It involves our relations with the enemy in the process of limiting war. In the interest of limiting war or of understanding limited war, it may be necessary to recognize that a distinction can exist between nuclear and other weapons even though the distinction is not physical but is psychic, perceptual, legalistic, or symbolic. That small-yield nuclears delivered with "pinpoint" accuracy are just a form of artillery, and consequently do not prejudice the issue of limits in war, is an argument based exclusively on an analysis of weapons effects, not on an analysis of the limiting process — of where limits originate in limited war, what makes them stable or unstable, what gives them authority, and what circumstances and modes of behavior are conducive to the finding and mutual recognition of limits. The premise of the "just-another-weapon" argument is that, if there is no compelling weapon-effects basis

for a distinction between nuclears and other weapons, there is no basis at all that is pertinent to the limiting process.

Is not the same point involved in discriminating among the users of weapons? There is no more difference between Russians and Chinese than there is between nuclear and other weapons; similarly for the difference between Chinese and North Koreans, or between Americans and Nationalist Chinese, British and Jordanians, Egyptians and Algerians. Yet nationality has been an important distinction in the process of limiting war or destroying its limits. Similarly, there is little difference between the terrain a hundred miles north of the Soviet-Iranian border and the terrain a hundred miles south, or what lies above the Yalu and below it, or the two sides of the Greek-Yugoslav border. Yet boundaries like these play an important role in the limiting process, quite aside from any physical difficulty in the crossing of rivers or the scaling of mountains that happen to coincide with them.

One could reply that these are "legal" distinctions and that legal distinctions are real ones while those between nuclear and other weapons are fictitious. But they are not really legal; they are "legalistic." There is no legal authority that forces the participants in limited war to recognize political boundaries or nationalities; the Russians are not legally obliged to treat a modest penetration of their border as a qualitative change in the war — as a dramatic act discontinuous with action up to their border. The Chinese were not legally obliged to retaliate (rather than just to resist) if we deliberately crossed the Yalu River; they did not lose any legal right to deny trespass by admitting occasional thoroughfare. We are not legally obliged to take cognizance of Russian pilots if they participate in a limited war, or Russian "volunteers" in a Near Eastern ground army fighting against our side. The inhibition on the penetration of a border, or on the introduction of a new nationality into the conflict, is like that on the introduction of a nuclear weapon; it is the risk of enemy response. And an important determinant of enemy response is his appreciation of what he has tacitly acquiesced in if he fails to respond, or makes only an incremental response, to our symbolically discontinuous act.

What makes the Soviet or Chinese border a pertinent or compelling place to draw a line in the event of war in that area is principally that there is usually no other plausible line to draw. For Western troops to cross the Russian border is to challenge — not physically but symbolically — the territorial integrity of the USSR, and to demonstrate or at least to imply an intention to proceed. Unless one can find some "obvious" limit inside that border, such that it would be clear to the Russians where we intended to stop in the event that we cross the border, and such that it would be obvious to us that there was a limit to how far the Russians would let us advance if we did cross it and that the Russians knew that we knew it, there is just no other stopping place that can be tacitly acknowledged by both sides. Under the circumstances for the USSR to accept the penetration of that border without a dramatic retaliation of some sort would be to admit that Soviet territory is fair game for a gradually expanding war. The political boundary is therefore *useful* as a stopping place, not legally mandatory; it is useful to *both* sides in default of any plainly recognizable alternative, since both sides have an interest in finding some limit. The border has a *uniqueness* that makes it a plausible limit. It is one of the few lines — perhaps the only line, but certainly one of the few — that one could draw in the region that could be tacitly recognized by both sides as the "obvious" geographical limit that both sides might observe. It has a compelling *power of suggestion,* a claim to attention, the denial of which might seem — in default of any plainly recognizable alternative — to be a denial of any limitation.

But, if political-boundary and nationality considerations still seem to be legal, and therefore real, consider some other distinctions that are significant in the limiting process. We provided much equipment but no manpower to the war in Indochina; we provided equipment, leadership, and advice to the Greek troops during the guerrilla war, but no combat troops. We provide direct naval support to the Nationalist Chinese in the Straits of Formosa. It has been thought that we might have given air support to the French and Vietnamese in Indochina, without appearing to the Chinese and Russians to be as "involved" as if we had put ground forces in.

An economist can argue — with the same persuasiveness as those who argue that "pinpoint"-delivered small-yield weapons are just another form of artillery — that equipment and manpower are fungible resources in a military campaign, that air intervention is not "really" different from ground intervention, that military intellect is as important as leg muscle for troops that lack leadership and planning skill. The controversy about redefinition of service functions in the light of modern weapons, and about the usefulness of defining military-service functions in terms of the means of locomotion, suggests that an air-ground distinction or a naval-ground distinction rests on nothing but tradition. But the point of all this is that, in limiting war, tradition matters.

In fact, what we are dealing with in the analysis of limited war is tradition. We are dealing with precedent, convention, and the force of suggestion. We are dealing with the theory of unwritten law — with conventions whose sanction in the aggregate is the need for mutual forebearance to avoid mutual destruction, and whose sanction in each individual case is the risk that to breach a rule may collapse it and that to collapse it may lead to a jointly less favorable limit or to none at all, and may further weaken the yet unbroken rules by providing evidence that their "authority" cannot be taken for granted.

What makes atomic weapons different is a powerful tradition that they *are* different. The reason — in answer to the usual rhetorical question — why we do not ban bows and arrows on the grounds that they too, like nuclear weapons, kill and maim people, is that there is a tradition for the use of bows and arrows, a jointly recognized expectation that they will be used if it is expedient to use them. There is no such tradition for the use of atomic weapons. There is instead a tradition for their nonuse — a jointly recognized expectation that they may not be used in spite of declarations of readiness to use them, even in spite of tactical advantages in their use.

Traditions or conventions are not simply an analogy for limits in war, or a curious aspect of them; tradition or precedent or convention is the essence of the limits. The fundamental characteristic of any limit in a limited war is the psychic, intellectual,

or social characteristic of being mutually recognized by both sides as having some kind of authority, the authority deriving mainly from the sheer perception of mutual acknowledgement, of a "tacit bargain." And a particular limit gains in authority from the lack of confidence that either side may have in what alternative limits may be found if the limit is not adhered to. The rationale behind the limit is legalistic and casuistic, not legal, moral, or physical. The limits may correspond to legal and physical differences or to moral distinctions; indeed, they usually have to correspond to something that gives them a unique and qualitative character and that provides some focus for expectations to converge on. But the authority is in the expectations themselves, and not in the thing that expectations have attached themselves to.

Whether limits on the use of atomic weapons, other than the particular limit of no use at all, can be defined in a plausible way is made more dubious, not less so, by the increasingly versatile character of atomic weapons. It is now widely recognized that there is a rather continuous gradation in the possible sizes of atomic-weapon effects, a rather continuous variation in the forms in which they can be used, in the means of conveyance, in the targets they can be used on, and so forth. There seems consequently to be no "natural" break between certain limited uses and others. If we ask, then, where we might draw a line if we wished to limit somehow the size of the weapons, the means of conveyance, the situations in which or the targets on which they can be used, the answer is that we are — in a purely technical sense — free to draw a line anywhere we please. There is no cogent reason for drawing it at any one particular gradation rather than another. But that is precisely why it is hard to find a rationale for any particular line. There is no degree of use, or size of weapon, or number of miles, that is so much more plausible than other degrees, sizes, or distances that it provides a focal point for both sides' expectations. Legalistic limits have to be qualitative and discrete, rather than quantitative and continuous. This is not just a matter of making violations easy to recognize, or of making adherence easy to enforce on one's own com-

manders; it concerns the need of any stable limit to have an evident symbolic character, such that to breach it is an overt and dramatic act that exposes both sides to the danger that alternative limits will not easily be found.

The need for qualitatively distinguishable limits that enjoy some kind of uniqueness is especially enhanced by the fact that limits are generally found by a process of tacit maneuver and negotiation. They are jockeyed for, rather than negotiated explicitly. But if the two sides must strike a "bargain" without explicit communication, the particular limit has to have some quality that distinguishes it from the continuum of possible alternatives; otherwise there is little basis for the confidence of each side that the other acknowledges the same limit. Even a parallel of latitude, or an international date line, or the north pole, may have this quality when no other natural, plausible, "obvious" point or line is available for expectations to converge on.

A test of this point with respect to atomic weapons might be to pose the following problem.[1] Let any of us try to cooperate for a prize: we are to sit down right now, separately and without any prior arrangements, and write out a proposed limitation on the use of nuclear weapons, in as little or as great detail as we please, allowing ourselves limitations of any description that appeals to us — size of weapons, use of weapons, who gets to use them, what rate or frequency of use, clean versus dirty, offensive versus defensive use, tactical versus strategic, on or not on cities, with or without warning — to see whether we can all write the *same* specification of limit. If we are in perfect agreement on the limits we specify, we get a prize; if our limits are different, we get no prize. We are doing this only for the sake of the prize, to see whether we can in fact agree tacitly on a statement of limits, and to see — for those of us who do manage to coordinate our proposals tacitly — what kinds of limits appear to be susceptible of tacit joint recognition. We are permitted the extremes of no limits at all on the one hand, or no atomic weapons at all on the other, and any gradation or variation defined in any way we please.

My argument is that there are particular limits — simple, dis-

[1] Compare Chapter 3, especially pp. 58–67.

crete, qualitative, "obvious" limits — that are conducive to a concerted choice; those who specify other kinds of limits, I predict, can find few partners or none at all whose limits coincide with theirs. (Since our object is to agree, we are to take no consolation in the other virtues of our proposed limits; in this exercise the main consideration in chosing any particular limits is the likelihood that if we chose those limits in an effort to coincide exactly with the limits of the others, knowing that they were trying to coordinate theirs with ours, we would succeed.)

I do not allege that this exercise proves what kinds of limits are capable of possessing stability and authority. It does demonstrate that certain characteristics of limits, particularly their simplicity, uniqueness, discreteness, susceptibility of qualitative definition, and so forth, can be given an objective meaning, one that is at least pertinent to the process of tacit negotiation. It suggests that certain kinds of limits are capable of being jointly expected by both sides, of focusing expectations and being recognized as qualitatively distinct from the continuum of possible alternatives.

The first conclusion to be drawn from this line of argument is that there is a distinction between nuclear and nonnuclear weapons, a distinction relevant to the process of limiting war. It is a distinction that to some extent we can strengthen or weaken, clarify or blur. We can strengthen the tradition, and enhance the symbolic significance of this distinction, by talking and acting in a way that is dramatically consistent with it; we can erode the distinction — but not readily destroy it — by acting as though we do not believe in it, by emphasizing the "just-another-weapon" argument and by making it evident that we in fact have little compunction about using nuclears. Which policy we should follow depends on whether we consider the distinction between nuclear and other weapons to be an asset that we share with the USSR, a useful distinction, a tradition that helps to minimize violence — or instead a nuisance, a propaganda liability, a diplomatic obstruction, and an inhibition to our decisive action and delegation of authority. Those who believe that atomic weapons ought to be used at the earliest convenience,

or whenever military expedience demands, should nevertheless recognize the distinction that exists so that we can take action to erode the distinction during the interim.

This is not just a matter of what the Asian neutrals or our European allies feel about the distinction. It concerns a relation between us and the Russians — an understanding that may exist between us whether we like it or not. It has to do with whether the Russians think we share with them a tacit expectation that there is a limit against the use of nuclear weapons. In the interest of limiting war, we should want the Russians or the Chinese not to believe that our initial use of atomic weapons in a local war were a challenge to the whole idea of limitations, a declaration that we would not be bound by any kinds of limits. We should want them to interpret our use of nuclear weapons as consistent with the concept of limited war and consistent with our willingness to collaborate tacitly in the discovery and recognition of limits; we should want our use of atomic weapons not to be charged with excessive symbolic content. So, if I am right that a distinction does exist in the sense pertinent to the limiting of war, and if nevertheless we want maximum freedom to use atomic weapons, we ought in the interest of limiting war to destroy or to erode the distinction as best we can. (For example, a deliberate program for early and extensive use of "nuclear dynamite" in earth-moving projects, especially in underdeveloped countries, might help to erode the distinction; the same might be true of a program for training friendly troops in underdeveloped countries in how to survive nuclear weapons explosions, using some actual weapons for the purpose in their own country.) If on the contrary we wish to enhance the tacit understanding we have with our enemies that nuclears are a class apart and subject to certain reservations, agreement on nuclear test suspension (or even just extensive discussion of such an agreement) will probably contribute to the purpose.[2]

A second conclusion is that the principal inhibition on the use of atomic weapons in limited war may disappear with their first

[2] On the symbolic significance of a test agreement, see Henry A. Kissinger, "Nuclear Testing and the Problem of Peace," *Foreign Affairs*, 37:1–18 (Oct. 1958), especially pp. 12–13.

use. It is difficult to imagine that the tacit agreement that nuclear weapons are different would be as powerfully present on the occasion of the *next* limited war after they had already been used in one. We can probably not, therefore, ignore the distinction and use nuclears in a particular war where their use might be of advantage to us and *subsequently* rely on the distinction in the hope that we and the enemy might both abstain. One potential limitation of war will be substantially discredited for all time if we shatter the tradition and create a contrary precedent. (There may also be some limits or sanctuary concepts that we take for granted that should be reexamined to see whether they were originally by-products of the assumed nuclear ban and might disappear with it. We may want to look again at the role of naval vessels, for example, partly to anticipate enemy treatment of them, partly to avoid misinterpreting enemy intentions if he treats them differently after nuclears are brought into play.)

A third conclusion is that on the occasion of their first use we should perhaps be at least as concerned with the patterns and precedents that we establish, and with the "nuclear role" that we adopt, as with the original objectives of the limited war. For example, if nuclear weapons were used in defense of Quemoy, we probably ought to be much less concerned about the outcome on Quemoy than about the character of the nuclear exchange, the precedents that it establishes, the role we manage to assume for ourselves, and the role the enemy assumes in the process. We shall be not only using them *ad hoc* for the little war in question, but importantly shaping the limited nuclear wars to come. (When a boy pulls a switch-blade knife on his teacher, the teacher is likely to feel, whatever the point at issue originally was, that the overriding policy question now is his behavior in the face of a switch-blade challenge.)

Fourth, we should recognize that — at least on the first occasion when nuclear weapons are used in limited war — the enemy too will really be engaged in at least two different kinds of limited-war activity at the same time. One will be the limited struggle over the original objectives; the second will be the tacit negotiation or gamesmanship over the role of nuclear weapons themselves. To illustrate, we might in connection with Quemoy decide to use

nuclear weapons; ordinarily it would be supposed that we should do this only if it were quite necessary to the defense of Quemoy, and that we should use them in a manner that achieves our Quemoy objectives. But, in considering whether the Chinese or Russians would use them in return, we should perhaps not worry mainly about what they think their use of nuclear weapons would do for the invasion of Quemoy. Much more important to them, it seems, would be the nature of their "response" to our nuclear initiative. They would be interested in not assuming a submissive role, but in demanding a kind of "parity" if not dominance in their own nuclear role. And, unless we are ready for some kind of decisive showdown in which we either win all or lose all, we must be as willing to "negotiate" (by our actions) for limited objectives in terms of nuclear dominance, traditions and precedents of nuclear use, and the "rules" we jointly create for future wars, as for any other types of objectives in limited war.

APPENDIX B

FOR THE ABANDONMENT OF SYMMETRY IN GAME THEORY

The first part of this appendix argues that the pure "moveless" bargaining game analyzed by Nash, Harsanyi, Luce and Raiffa, and others,[1] may not exist or, if it does, is of a different character from what has been generally supposed; the point of departure for this argument is the operational meaning of *agreement*, a concept that is almost invariably left undefined. The second part of the paper argues that symmetry in the solution of bargaining games cannot be supported on the notion of "rational expectations"; the point of departure for this argument is the operational identification of irrational expectations.

A nontacit ("cooperative") nonzero-sum game — a bargaining game — is not *defined* by its payoff matrix; the operations by which choices are made must still be specified. Commonly these operations are sketched in by reference to the notion of "binding agreements" and the notion of free communication in the process of reaching agreement. Thus to say that two players may divide $100 as soon as they can agree on how to divide it, and that they may discuss the matter fully with each other, is generally considered sufficient to define a game.[2]

[1] John F. Nash, "The Bargaining Problem," *Econometrica*, 18:155–162 (April 1950), and "Two-Person Cooperative Games," *Econometrica*, 21:128–140 (January 1953); John Harsanyi, "Approaches to the Bargaining Problem Before and After the Theory of Games: a Critical Discussion of Zeuthen's, Hicks', and Nash's Theories," *Econometrica*, 24:144–157 (April 1956); R. Duncan Luce and Howard Raiffa, *Games and Decisions* (New York, 1957), pp. 114ff.

[2] Luce and Raiffa, in effect, *define* cooperative two-person games by reference to a payoff matrix and the following three stipulations. (1) All preplay messages formulated by one player are transmitted without distortion to the

A game of this sort is symmetrical in its move structure, even though it may be asymmetrical in the configuration of payoffs. The two players have identical privileges of communication, of refusing offers, and of reaching agreement. If instead of dividing $100 the players are to agree on values X and Y contained within a boundary, the payoff function may not be symmetrical but the move structure is. Harsanyi, to emphasize this, has even added explicitly the postulate of symmetrical moves: "The bargaining parties follow identical (symmetric) rules of behaviour (whether because they follow the same principles of rational behaviour or because they are subject to the same psychological laws)." [3]

What I want to do is to look at this notion of "agreement" on the assumption of *perfect symmetry in the move structure of the game*, paying close attention to the "legal details" of the bargaining process. We must also look at the meaning of "nonagreement." Since any well-defined game must have some rule for its own termination, let us look at the rules for termination first.[4]

If we are to avoid adding a whole new dimension to our payoff matrix, in the form of discount rates, we must suppose that the game is terminated soon enough so that nothing like the interest rate enters the picture. We do not want to have to consider the *time* at which agreement is reached, in addition to the agreement itself. This is more than a matter of convenience; the game ceases to be "moveless," except in very special cases, unless we make this stipulation. For, if the players' time preferences take any shape except that of a continuously uniform discount rate, the game itself changes with the passage of time and a player can, in effect, change the game itself by failing to reach agreement. The notion of a continuously uniform discount rate is probably far too special to treat as a *necessary* condition, and anyway has not been made

---

other player. (2) All agreements are binding, and they are enforceable by the rules of the game. (3) A player's evaluation of the outcomes of the game are not disturbed by these preplay negotiations. *Games and Decisions*, p. 114.

[3] John Harsanyi, "Approaches to the Bargaining Problem Before and After the Theory of Games . . . ," *Econometrica*, 24:149 (April 1956).

[4] The model discussed here is quite abstract, artificial, and unrealistic; but it does have the advantage of helping to test whether *even in an artificially abstract model* it is fruitful to postulate perfect symmetry in the move structure and to treat asymmetry as a special case, symmetry as the more general case.

an explicit postulate in the models under examination; so we must assume that the game is somehow gotten over with.

Perhaps the simplest way to terminate the game is to have a bell ring at a time specified in advance. There are other ways, such as having the referee roll dice every few minutes, calling off the game whenever he rolls boxcars. (We might have the game terminate after a specified number of offers have been refused, but this would change the character of the game by making certain kinds of communication "real moves" that leave the game different from what it was before, and perforce lead us into such tactics as the exhaustion of offers.)

For simplicity, suppose that the game will be terminated at a time specified in advance to the players, and for convenience let us call the final moment "midnight." If agreement exists when the midnight bell rings, the players divide the gains in the way they have agreed; if no agreement exists, the players receive nothing.

Next, what do we mean by "agreement"? For simplicity, suppose that each player keeps (or may keep) his current "official" offer recorded in some manner that will be visible to the referee when the bell rings. Perhaps he keeps it written on a blackboard that the other player can see; perhaps he keeps it in a sealed envelope that is surrendered to the referee when the bell rings; perhaps he keeps it punched into a private keyboard that records his current offer in the referee's room. When the bell rings, the blackboard is photographed, the envelope surrendered, or the keyboard locked, so that the referee needs only to inspect the two "current" offers as they exist at midnight to see whether they are compatible or not. If they are compatible, the gains are divided in accordance with the "agreement"; if the two players have jointly claimed more than is available, "disagreement" exists and the players get nothing. (Defer, for a moment, ruling on what happens if the two players together have claimed less than the total available, whether they get as much as they have claimed or get nothing for lack of proper agreement. And, in what follows, it will not matter whether an exhaustive agreement reached before midnight — that is, compatibility of the current offers occurring before midnight — terminates the game.)

There are other ways of defining "agreement" in terms of the operations by which it is reached or recorded; but if we adhere to the notion of a *perfectly symmetrical move structure* they will generally, I think, have the property that I am trying to single out for attention. That property is this. There must be some minimum length of time that it takes a player to make, or to change, his current offer. (For simplicity again, let us suppose that the same operation either makes an offer or changes it, so that we may always assume that a "current offer" exists.) There must then be some critical moment in time, a finite period before the midnight bell rings, that is the last moment at which a player can begin the operations that record his final offer. That is, there is some last moment before the bell rings, beyond which it is too late to change one's existing offer. Under the rules of the game and the rationality postulate both players know this. And by the rule of symmetry this moment must be the same for both players.

From this follows the significant feature. The last offer that it is mechanically and legally possible for a player to make is one that he necessarily makes without knowing what the other player's final offer is going to be; and the last offer that a player can make is one that the other player cannot possibly respond to in the course of the game. Prior to that penultimate moment, no offer has any finality; and at that last moment players either change or do not change their current offers, and whatever they do is done in complete ignorance of what each other is doing, and is final.[5]

This must be true. If either could get a glimpse of the other's final offer in time to do anything about it, or if either could give the other a glimpse of his own final offer in time for the other to respond, it is not — and is known to be not — a final offer.[6]

But now we have reached an important conclusion about the

[5] Incidentally, the argument is unaffected by supposing that a player can change his offer "instantaneously" as long as we keep the symmetrical rule that both can do it "equally instantaneously" as the final bell rings.

[6] There is a mechanical assumption here that in the process of making a new offer one can stop and start over. The case is slightly more complicated if an offer started one and one-half minutes before midnight is necessarily the last offer because the process cannot be started again until a minute has passed and by then the critical point has been passed. This case will be looked at again below.

perfectly move-symmetrical bargaining game. It is that it necessarily gives way, at some definite penultimate moment, to a *tacit* (noncooperative) bargaining game. And each player knows this

The most informative way to characterize the game, then, is not that the players must reach overt agreement by the time the final bell rings or forego the rewards altogether. It is that they must reach overt agreement by a particular (and well-identified) penultimate moment — when the "warning bell" rings — *or else play the tacit variant of the same game.*

Each player must be assumed to know this and may, if he wishes, by simply avoiding overt agreement, elect to play the tacit game instead. So, if we assume (for the moment) that the tacit game has a clearly recognized solution, and that the solution is *efficient*, each player has a pure minimax behavior strategy during the earlier stage. Either can enforce this tacit solution by abstaining from agreement until the warning-bell rings; neither can achieve anything better from a rational opponent by verbal bargaining.

From this it follows that the solution of the cooperative game must be identical with that of the corresponding tacit game (if the latter has a predictable and efficient solution). It must be, because the tacit game comes as an inevitable, mechanical sequel to the cooperative game.

At this point it looks as though the cooperative feature of the game is irrelevant. The players really need not show up until 11:59; in fact they do not need to show up at all. The preplay communication and ability to reach binding agreements, which were intended to characterize the game, prove to be irrelevant; the cooperative game as a distinct game from the tacit game does not exist.[7]

But this conclusion is unwarranted. First, a tacit game may not

[7] In his 1953 article, "Two-person Cooperative Games," Nash presents a model that is explicitly tacit in its final stage. The model's relation to the cooperative game was heuristic: it was to help to discover what might constitute "rational expectations" (and hence the indicated rational outcome) in the corresponding cooperative game. The argument of the present paper is that the relation is likely to be mechanical rather than intellectual if a symmetrical move structure is strictly adhered to, and that with strict symmetry it is difficult, perhaps impossible, to define the corresponding nontacit game that was the ultimate subject of study.

have a confidently predicted efficient solution.[8] More than that, certain details of the cooperative game that might have seemed to be innocuous from the point of view of explicit negotiation may affect the character of the tacit game; similarly, preplay communication that has no binding effect on the players themselves may also affect the character of the tacit game. For an example, consider the following variant of the cooperative game.

Instead of saying that the players may divide a set of rewards if they can reach agreement on an exhaustive division, let us say that the players may divide a set of rewards *to the extent* that they have reached agreement on a division; they may divide such portion of the available rewards as they have already reached agreement on by the time the bell rings. If, for example, there are one hundred individual objects and the players have reached agreement on how to divide eighty of them when the bell rings, the twenty items in dispute revert to the house while the eighty on which agreement was reached will be divided in accordance with the agreement.[9]

[8] It should be emphasized that bargaining-game solutions that (like the Nash and Harsanyi solutions) depend on a clearly recognized zero point — that is, on an unambiguous outcome that reigns in the absence of overt agreement — cannot necessarily be applied to a cooperative game that is based on a matrix of choices. A matrix (unless perhaps all payoffs are zero except in the diagonal) does not have a zero point defined by the rules. There is consequently no "normal form" consisting of a convex region and associated zero point unless there is available a fully adequate theory that "solves" the tacit game (and does so in a manner that the players can take for granted) One may, following Luce and Raiffa (for example, page 137) take the players' "security levels" (maximin values) as the zero point; but this is either arbitrary or based on the hypothesis that, left to themselves, the players could succeed in doing no better than this in the tacit game. The latter hypothesis especially where there are pure-strategy efficient points (as in Braithwaite's game, and as in the Luce-Raiffa matrix discussed in note 18 below), is a weak hypothesis that can be empirically refuted; it assumes that rational players are incapable of correlating strategies without communicating, while in fact this is something they often can do even in the face of conflicting preferences (This point is taken up again in note 18.) The potential ambiguity of the zero point is the issue between Harvey Wagner and John Harsanyi in the former's "Rejoinder on the Bargaining Problem," *Southern Economic Journal*, 24 480-482 (April 1958).

[9] In the case of a single divisible object like money, the corresponding rule might be that they divide the money in accordance with their offers after the house has removed the "overlap." Each player obtains as much as the other implicitly accords him; if one is demanding 65 percent of the money at the

Now, in the explicit-bargaining (cooperative) case, if we had already concluded there was an efficient solution to this game — that is, that the players would in fact reach an exhaustive agreement — we should probably have considered this reformulation of the problem inconsequential. The reformulation says, in effect, only that bargaining should take the form of each player's writing down the totality of his claim and that concessions shall take the form of each player's deleting items from his list of claims, with full agreement being reached when no more items are in conflict on the lists of claims. But, when we look at the tacit case, the game is drastically altered by this reformulation. The tacit game now has a perverse incentive structure. There is no rational reason for either player to demand less than the whole of the available reward; each knows this and knows that the other knows it. There is no incentive to reduce one's claim because any residual dispute costs the player no more than he would lose if he reduced his claim to eliminate the dispute. The single equilibrium point yields zero for both players. Thus the variant game, which seemed to differ inconsequentially, is drastically different from the original game; but it does not appear so until we have identified the terminal tacit game as a dominating influence.[10]

To take another example, suppose there are 100 individual objects to be divided and that, although they are fungible as far as value is concerned, the agreement must specify precisely which *individual items* go to which individual players. If the rules require that full and exhaustive agreement be reached, then in the tacit game the players are dependent on their ability not only to divide the total value of the objects in coordinated fashion but to

end of the game, and the other 55 percent, the second has been accorded 35 percent and the first 45 percent; these amounts are outside the range of dispute and constitute the "agreement."

[10] It might seem that we can draw a by-product from the analysis here, namely, the observation that in order to set up a "truly" cooperative (non-tacit) game, the legal definition of agreement must be such as to make the ultimate tacit game perverse, so that the players must reach binding agreement before the warning bell or suffer complete loss. But there is still a problem. The players themselves must now define "agreement" for purposes of their own agreement prior to the final bell. If it is like our earlier definition, all they accomplish is to make the perverse cooperative game into a benign one, one minute shorter, which is equivalent to a tacit game two minutes shorter than the original.

sort out the 100 individual objects into two piles in identical fashion. If, then, one of the players has demanded *specific* items worth 80 percent of the total and the other player has refused, the former has an advantage in the tacit game. The only extant proposal for dividing the 100 objects is the one player's specification of 80 that would satisfy him; the chances of their concerting identically on any other division of the 100 objects, equal or unequal between them, may be so small that they are forced for the sake of agreement into accepting the only extant proposal in spite of its bias. Thus preplay communication has tactical significance in that it can affect the means of coordination once the tacit stage of the game has been reached.

If now, in considering the tactical implications of this last point, we insist on a rule of symmetrical behavior, we must conclude that if either player opened his mouth to drown out what the other was about to say, he would always find the other player also with his mouth open, both knowing that if either spoke the other would be found to be speaking, neither able to hear the other, and so on. In other words, the assumption of complete symmetry of behavior as a recognized foregone conclusion seems to preclude the very kind of action that might have seemed to enrich the game at the stage of preplay communication.

But by now we have certainly pressed the perfect move-symmetrical game as far as is worthwhile.[11] We could go on to ana-

[11] One detail may be worth pursuing, in line with an earlier footnote. Suppose that it takes one minute to make or change an offer and (in contrast to the earlier version) that the process of recording a new offer, once started, cannot be stopped before it is completed. Under this procedure, any offer initiated during the next to last minute of the game is one's final offer. If this final offer *cannot* be communicated to the other player before the expiration of the minute, the game is essentially the same as before; "simultaneous" now means within a minute of each other for practical purposes, and again neither can see the other's final offer as he initiates his own, no matter what time during the final minute the offers are initiated. But suppose one punches his offer into a visible board which remains locked for one minute while the offer is recorded, so that the other player can see one's offer in a few seconds although one cannot initiate a change until the minute's delay is up. (And suppose that neither can make himself visibly incapable of seeing the other's offer once it is so recorded.) In this case, if the two offers during that final minute are not simultaneous, the player who moves second makes his final offer in full knowledge of the other's; and since his only chance of winning anything is to accept it, he must accept whatever the other has offered. Thus "second move" loses

lyze this game in more detail, considering such things as alternative ways of terminating the game or of defining "agreement," and so forth. It seems more worthwhile, however, to raise at this point the question of whether the perfectly "moveless" or "move-symmetrical" game is a profitable one to study. Is the nondiscriminatory, move-symmetrical game a "general" game, one that gets away from "special cases"? Or is it a special, limiting case in which the most interesting aspects of the cooperative game have vanished?

It should be emphasized that the fruitful alternative to symmetry is not the assumption of asymmetry, but just *nonsymmetry*, admitting both symmetry and asymmetry as possibilities without being committed to either as a foregone conclusion.

An illustration may help. Suppose we were to analyze the game in which there is $100 at the end of the road for the player who can get there first. This game of skill is not hard to analyze: the money goes to the fastest, barring accidents and random elements. We can predict rational behavior (running) and the outcome (money to the fastest). Ties will occasionally occur; but they will occur at the end of a race and will not be taken for granted at the outset. We need an auxiliary rule to cover ties, but it need not dominate either the game or the analysis.

Consider the same game played in a population in which everybody can run exactly as fast as anybody else, and everybody knows it. Now what happens? Every race ends in a tie, so the

---

if the first mover knows that the other is waiting. We now have a game that can be characterized as follows: the players dally around for 23 hours 58 minutes and then play a game lasting one minute, this game allowing each player one and only one offer which he can make at any time during the minute. This game offers, in effect, three strategies to a player, namely, (1) assume the other will wait, and demand 99 per cent; (2) assume both will make simultaneous offers, and demand whatever is indicated by the tacit game; (3) wait. If both wait, the game is still to be played. If there is a finite number of potential waits, we have strategies of wait-once-then-demand-99-per-cent, wait-once-demand-tacit-solution; wait-twice-demand-99-per-cent, wait-twice-demand-tacit-solution; and so on. This game (the "tacit supergame" consisting of all strategies for playing the one-minute game) is then *the* game; and it has, if we wish to accept it, its own "solution in the strict sense" which consists of all strategies (all lengths of waits) that end in demands that correspond to the solution of the tacit game. (For the definition of a solution in the strict sense in a tacit two-person game, see Appendix C.)

auxiliary rule is all that matters. But since a tie is a foregone conclusion, why would they bother to run?

The perfectly move–symmetrical cooperative game seems a little like that foot race. Bargaining in the one case is as unavailing as leg-work in the other; every player knows in advance that all moves and tactics are foredoomed to neutralization by the symmetrical potentialities available to his opponent. The interesting elements that we might inject in the bargaining game are meaningless if perfect symmetry, and its acceptance as inevitable by both players, are imposed on the game by its definition.

What should we add to the game to enrich it if the assumption of symmetry is dropped? There are many "moves" that are often available, but not necessarily equally available to both players, in actual game situation. "Moves" would include commitments, threats, promises; tampering with the communication system; invocation of penalties on promises, commitments, and threats; conveyance of true information, self-identification; and the injection of contextual detail that may constrain expectations, particularly when communication is incomplete. Such "moves" were discussed in detail in Chapters 2–5.

To illustrate, suppose in the earlier cooperative game there is a turnstile that permits a player to leave but not to return; his current offer as he goes through the turnstile remains on the books until the bell rings. Now we have a means by which a player can make a "final" offer, a "commitment"; whoever can record an offer favorable to himself and known to the other, and leave the room, has the winning tactic. Of course it may win for either of them; but this may mean that we end up with something like a foot race, and the one closest to the turnstile wins. By analyzing the tactic, and its institutional or physical arrangements, we may determine who can make first use of it.

We have not, it should be noted, converted the game of strategy into a game of skill by letting them race for the turnstile. It remains true that one wins when he gets to the turnstile first only through the other's cooperation, only by constraining the other player's choice of strategy. He does not win legally or physically by going through the turnstile; he wins *strategically*. He makes the other player choose in his favor. It is a tactic in a game of

strategy, even though the *use* of it may depend on skill or locational advantage.

We can even put a certain kind of symmetry into the game now, without destroying it; we can flip a coin to see who is nearest the turnstile when the game begins, or let the players be similarly located and similar of speed but with random elements to determine who gets to the turnstile first. Though the *game* is now *nondiscriminatory*, the *outcome* would still be *asymmetrical* because each player has an incentive to run to the turnstile, leaving behind a standing offer in his own favor.[12]

We can include some risk of "tie," especially if there are two turnstiles and the players might go through them simultaneously. This constitutes "symmetry" as an interesting possibility, but not as a foregone conclusion; stalemate and the anticipation of it become interesting possibilities if the actions and information structure are in fact conducive to ties. But, with nonsymmetry as our philosophy, we do not need to be obsessed with the possibility of ties.

Again, if one player can make an offer and destroy communication, he may thereby win the ensuing tacit game by having provided the only extant offer that both players can converge on when they badly need to concert their choices later during the final tacit stage. To be sure, we can consider what happens when identical capacities for destruction of communication are present, and both players must recognize that they may simultaneously destroy communication without getting messages across; but this interesting case seems to be a special one, not the general case.

In summary, the perfectly "moveless" or "move-symmetrical" cooperative game is not a fruitful general case, but a limiting case that may degenerate into an ordinary tacit game. The cooperative game is rich and meaningful when "moves" are admitted; and much of the significance of the moves will vanish if complete symmetry in their availability to the players is stamped into the definition of the game. It is the moves that are interesting, not the

[12] It could be argued at this point that the expected value of the game is still symmetrically divided between the players, and that the analyst may consequently still view the game as symmetrical in terms of average outcomes. But if he does so he commits himself to a minimum of insight into the game and the way the game will be played.

game without moves; and it is the potential asymmetry of the moves that makes them most interesting.

Symmetry is not only commonly imposed on the move-structure of games but adduced as a plausible characteristic of the solution of the game or of the rational behavior with which the solution must be consistent. Nash's theory of the two-person cooperative game explicitly postulates symmetry, as does Harsanyi's. The symmetry postulate is certainly expedient; it often permits one to find a "solution" to a game and to stay — if he wishes to — within the realm of mathematics. There are few similarly potent concepts that compete with it as bases for solving a game. But the justification for the symmetry postulate has not been just that it leads to nice results; it has been justified on grounds that the contradiction of symmetry would tend to contradict the rationality of the two players. This is the underpinning that I want to attack.

What I am going to argue is that, though symmetry is consistent with the rationality of the players, it cannot be demonstrated that asymmetry is inconsistent with their rationality, while the inclusion of symmetry in the *definition* of rationality begs the question. I then want to offer what I think is *an* argument in favor of symmetrical solutions, an argument that tends to make symmetry but one of many potential influences on the outcome with no prima facie claim to pre-eminence.

Explicit statements of the relation between symmetry and rationality have been given by John Harsanyi. He says, "The bargaining problem has an obvious determinate solution in at least one special case: viz., in situations that are completely symmetric with respect to the two bargaining parties. In this case it is natural to assume that the two parties will tend to share the net gain equally since neither would be prepared to grant the other better terms than the latter would grant him." [13] In a later paper he refers to the symmetry axiom as the "fundamental postulate"

[13] Harsanyi, 147. He goes on to say, "For instance, everybody will expect that two duopolists with the same cost functions, size, market conditions, capital resources, personalities, etc., will reach an agreement giving equal profits to each of them."

and says, "Intuitively the assumption underlying this axiom is that a rational bargainer will not expect a rational opponent to grant him larger concessions than he would make himself under similar conditions." [14]

Now this intuitive formulation involves two postulates. First, that one bargainer will not concede more than he would expect to get if he himself were in the other's position. Second, that the only basis for his expectation of what he would concede if he were in the other's position is his perception of symmetry.

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

[14] The full quotation deserves to be given: "What the Zeuthen-Nash theory of bargaining essentially proposes to do is to specify what are the expectations that two rational bargainers can consistently entertain as to each other's bargaining strategies if they know each other's utility functions. The fundamental postulate of the theory is a symmetry axiom, which states that the functions defining the two parties' optimal strategies in terms of the data (or, equivalently, the functions defining the two parties' final payoffs) have the same mathematical form, except that, of course, the variables associated with the two parties have to be interchanged. Intuitively the assumption underlying this axiom is that a rational bargainer will not expect a rational opponent to grant him larger concessions than he would make himself under similar conditions." (Harsanyi, "Bargaining in Ignorance of the Opponent's Utility Function," Cowles Foundation Discussion Paper No. 46, December 11, 1957, quoted by permission of the author.)

and what they expect is that they both perceive the same solution.[15]

In these terms the first (explicit) part of the Harsanyi hypothesis might be rephrased: that there is, in any bargaining-game situation (with perfect information about utilities), a particular outcome such that a rational player on either side can recognize that any rational player on either side would recognize it as the indicated "solution." The second (implicit) part of the hypothesis is that the particular outcome so recognized is determined by mathematical symmetry. The first we might call the "rational-solution" postulate; it is the second that constitutes the "symmetry" postulate.

The question now is whether the symmetry postulate is *derived* from the players' rationality — the rationality of their expectations — or must rest on other grounds. If it rests on other grounds, what are they and how firm is the support?

To pursue the first question, whether symmetry can be de-

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

There is a qualification to this point. With full information about each other's value systems and a homogeneous set of gains to be divided, there may be an infinity of equivalent solutions, all yielding the same values to the two players, but no difficulty in agreeing on an arbitrary choice among this indifferent set. But tacit bargaining often requires a further degree of coordination, namely, a coordinated choice even among equivalent divisions of the gains. Negotiation over a boundary line in homogeneous territory is thus different from the simultaneous dispatch of troops to take up positions representing claims (as in Question 6 on page 62); such claims may overlap and cause trouble even though the terrain values claimed are consistent. Thus the coordination problem is different; and there is no *a priori* assurance that the solution to the tacit game (or to games with somewhat incomplete communication, information, and so forth) would be in the set of equivalent solutions to the fully explicit game.

duced from the rationality of the players' expectations, we can consider the rationality of the two players jointly and inquire whether a jointly expected nonsymmetrical outcome contradicts the rationality postulate. If two players confidently believe they share, and do share, the expectation of a particular outcome, and that outcome is not symmetrical in a mathematical sense, can we demonstrate that their expectations are irrational, and that the rationality postulate is contradicted? Specifically, suppose that two players may have $100 to divide as soon as they agree explicitly on how to divide it; and they quite readily agree that A shall have $80 and B shall have $20; and we know that dollar amounts in this particular case are proportionate to utilities, and the players do too. Can we demonstrate that the players have been irrational?

We must be careful not to make symmetry part of the *definition* of rationality; to do so would destroy the empirical relevance of the theory and simply make symmetry an independent axiom. We must have a plausible definition of rationality that does not mention symmetry and show that asymmetry in the bargaining expectations would be inconsistent with that definition. For our present purpose we must suppose that two players have picked $80 and $20 by agreement and see whether we can identify any kind of intellectual error, misguided expectations, or disorderly self-interest, on the part of one or both of them, in their failure to pick a symmetrical point.

Specifically, where is the "error" in B's concession of $80 to A? He expected — he may tell us, and suppose that we have means to check his veracity (a modest supposition if full information of utilities is already assumed!) — that A would "demand" $80; he expected A to expect to get $80; he knew that A knew that he, B, expected to yield $80 and be content with $20; he knew that A knew that he knew this; and so on. A expected to get $80, knew that B was psychologically ready because he, B, knew that A confidently expected B to be ready, and so on. That is, they both knew — they tell us — and both knew that both knew, that the outcome would ineluctably be $80 for A and $20 for B. Both were correct in every expectation. The expectations of each were internally consistent and consistent with the other's.

We may be mystified about *how* they reached such expectations; but the feat claims admiration as much as contempt. The "rational-solution" postulate is beautifully borne out; the game seems to have dictated a particular outcome that both players confidently perceived. If, at this point, we feel that we ourselves wouldn't have perceived the same outcome, we can conclude that one of four hypotheses is false: (1) the rational-solution postulate, (2) the rationality of A and B, (3) our own rationality, (4) the identity (in all essential respects) of the game that we introspectively play with the game that A and B have just played. But we cannot, on the evidence, declare the second to be the false one — the rationality of A and B.

Note that if B had insisted on $50, or if A had been content to demand $50, claiming to be rational and arguing in terms of confidence in a shared expectation of that outcome, both players would have been in "error" and we could not tell, on the evidence, which one was irrational or whether they both were. Unless we made symmetry the definition of rationality we could only conclude that at least one of the players was irrational or that the rational-solution postulate did not hold. What we have is at best a single *necessary* condition for the irrationality of both players jointly; we have no sufficient condition, and no necessary condition that can be applied to a single player.

Nor can we trip them up if we ask them how they arrived at their expectations. Any grounds that are consistent would do, since any grounds that each expects the other confidently to adopt are grounds that he cannot rationally eschew. Consistent stories are all they need; and if they say that a sign on the blackboard said A-$80, B-$20, or that they saw in a bulletin that two other players, named A′ and B′, split $80-$20, and that they confidently perceived that this was clear indication to both of them of what to expect — that this was the only "expectable" outcome — we cannot catch them in error and prove them irrational. They may be irrational; but the evidence will not show it.

There is, however, a basis for denying my present argument. Since I have not actually applied an independent test of rationality to two players, given them the game to play, and observed the 80:20 split that I just mentioned, but have only posed it as a

possibility to see whether it would imply irrationality *if* it occurred, one might object that it could not occur. And the argument would rest on the problem of coordination; it would run as follows.

If two players jointly expect *a priori* the same outcome, and confidently recognize it as their *common* expectation, they must have the intellectual power to pick a particular point in common. If the whole $100 can be divided to the nearest penny, there are 9,999 relevant divisions to consider, one of which would have to be picked simultaneously but separately by both players as their expectations of the outcome. But how can two people concert their selections of one item out of 9,999, in the sense that their expectations focus or converge on it, except with odds of 9,999 to 1 against them? The answer must be that they utilize some trick, or clue, or coordinating device that presents itself to them. They must, consciously or unconsciously, use a selection procedure that leads to unique results. There must be something about the point they pick that distinguishes it — if not in their conscious reasoning, at least in our conscious analysis — from the continuum of all possible alternatives.

Now, is it possible for two rational players, through anything other than sheer coincidence or magic, to focus their attention on the same particular outcome and each "rationally" be confident that the other is focussed on the same outcome with the same appreciation that it is mutually expected? And, if so, how can they?

The answer is that they can, as demonstrated in Chapter 3. They may use any means that is available: any clue, any suggestion, any rule of elimination that leads to an unambiguous choice or a high probability of concerted choice. And one of these rules, or clues, or suggestions, is mathematical symmetry.[16]

[16] The basic intellectual premise, or working hypothesis, for rational players in this game seems to be the premise that *some* rule must be used if success is to exceed coincidence, and that the best rule to be found, whatever its rationalization, is consequently a rational rule. This premise would support, for example, Nash's model that views an "unsmoothed" tacit game as the limit of a "smoothed" game as the smoothing approaches zero. While this view of the unsmoothed game is in no sense logically necessary, it is a powerfully suggestive one that can, in the absence of any better rationale for converging on a single point, command the attention of players in need of a common choice.

In a game that has absolutely no details but its mathematical structure, in which no inadvertent contextual matter can make itself appreciated by a player as something that the other can appreciate too, there may be nothing to work on but a continuum of numbers. And all the numbers can be sorted according to whether they correspond to symmetrical or asymmetrical divisions. If all numbers but one represent an asymmetrical split, then sheer mathematical symmetry is a sufficient rule and a supremely helpful one in concerting on a common choice. And it may be possible to set up a game in such sanitary fashion, suppressing the identity of players and all contextual details, that there is literally no other visible basis for concerting unless impurities creep in.[17]

In other words, mathematical symmetry may focus the expectations of two rational players because it does — granted the other assumed features of the game, like full information on each other's utility systems — provide one means of concerting expectations. Whether it is a potent means may depend on what alternatives are available.

That there are other means of concerting, including some that may substantially outweigh the notion of symmetry, seems amply demonstrated by the experiments in Chapter 3. So it is demonstrably possible to set up games in which mathematical sym-

---

The limiting process provides a clue for picking one of the infinitely many equilibrium points that actually exist in the unsmoothed game. Of course, the premise equally supports any other procedure that produces a candidate for election among the infinitely many potential choices.

[17] In this view, the theory of Nash (leading to the maximum-utility-product solution) is a response to the fact that even in the realm of mathematics there are offhand too many types of uniqueness or symmetry to provide an unambiguous rule for selection, hence a need to adduce plausible criteria (axioms) sufficient to yield an unambiguous selection. Braithwaite's theory can be characterized the same way. The fact that the two solutions conflict implies that mathematicians may not have a sufficiently common mathematical aesthetic to satisfy the first part of the Harsanyi postulate, that is, to coordinate their expectations on the same outcome. (R. B. Braithwaite, *Theory of Games as a Tool for the Moral Philosopher* [Cambridge, England, 1955]; Braithwaite's solution is described in Luce and Raiffa, *Games and Decisions*, 145ff.) Braithwaite's construction of the problem as a one-person arbitration problem, and Luce and Raiffa's reformulation of Nash's theory in terms of arbitration rather than strategy (pages 121–154), seem to emphasize that *intellectual coordination* is at the heart of the theory. A legalistic solution requires *some* rationalization of a unique outcome; pure casuistry is helpful if the alternative is vacuum.

metry does provide the focus for coordinated expectations, and demonstrably possible to set up games in which some other aspect of the game focusses expectations. (These other aspects are commonly not contained in the mathematical structure of the game but are part of the "topical content"; that is, they usually depend on the "labeling" of players and strategies, to use the term of Luce and Raiffa mentioned in Chapter 4.)

I have no basis for arguing with what force, or in what percentage of interesting games, mathematical symmetry does dominate "rational expectations." But I think that the status of the symmetry postulate is qualitatively changed by the admission that symmetry has competitors in the role of focussing expectations. For, if it were believed that rational players' expectations could be brought into consistency only by some mathematical property of the payoff function, then symmetry might seem to have undisputed claim, particularly if it is possible to find a unique definition of symmetry that meets certain attractive axioms. But if one has to admit that other things — things not necessarily part of the mathematical structure of the payoff function — can do what symmetry does, then there is no *a priori* reason to suppose that what symmetry does is 99 percent or 1 percent of the job. The appeal of symmetry is no longer mathematical, it is introspective; and further argument is limited to the personal appeal of particular focussing devices to the game theorist *as game player,* or else to empirical observation.

Thus a normative theory of games, a theory of strategy, depending on intellectual coordination, has a component that is inherently empirical; it depends on how people can coordinate their expectations. It depends therefore on skill and on context. The rational player must address himself to the empirical question of how, in the particular context of his own game, two rational players might achieve tacit coordination of choices, if he is to find in the game a basis for sharing an *a priori* expectation of the outcome with his partner. The identification of symmetry with rationality rests on the assumption that there are certain intellectual processes that rational players are incapable of, namely, concerting choices on the basis of anything other than mathematical symmetry, and that rational players should know this. It is an

empirical question whether rational players can actually do what such a theory denies they can do and should consequently ignore the strategic principles produced by such a theory.[18]

An introspective game, which could be submitted to experiment, may illustrate the point. Imagine a game's potential payoffs as consisting of all the points on or within some boundary in the upper-right quadrant relative to a pair of rectangular coordinates.

[18] It is interesting that in demanding a symmetrical solution to an ostensibly symmetrical tacit game, Luce and Raiffa dismiss the two most promising candidates. They consider (*Games and Decisions*, 90–94) a matrix,

	I	II
i	1 2	−1 −1
ii	−1 −1	2 1

and note that it has pure-strategy equilibrium points in the upper-left and lower-right corners. These are ruled out on grounds that "whatever rationalization I give for either i or ii there is, by the symmetry of the situation, a similar rationalization for player 2, and so it seems inevitable that we both lose." (I have substituted i and ii for their designations.) They then look at a pair of maximin strategies, which are unsatisfactory because they do not produce an equilibrium point, and a minimax strategy which they find even inferior. But the important question is whether players who are both rational and imaginative are quite as impotent as Luce and Raiffa insist. Can players correlate strategies without communicating? This an empirical question; the experiments of Chapter 3 give an affirmative answer, or at least indicate that in particular cases the answer may be yes. Offhand it may seem hard for them to concert on a nonsymmetrical pair of strategies. But much the hardest part is just recognizing that they have to; the question of how to do it then becomes a practical matter. They must jointly and tacitly find a clue to the concerting of their choice. Of course, a nonsymmetrical solution in the above matrix is a discriminatory one; it quite arbitrarily condemns one of the players to a smaller gain than the other for reasons that may seem purely accidental or incidental. But we have to suppose that a rational player can discipline himself to accept the lesser share if the clue points that way. Only a discriminatory clue can point to a concerted choice; to deny the discrimination is to deny the premise that a clue can be jointly found and jointly acted on in the interest of an outcome that is jointly far superior to any symmetrical outcome. Luce and Raiffa conclude their discussion of this particular game with the remark that "although this seemingly innocuous game possesses some symmetries it is difficult to see how to exploit them." But the real key to this seemingly innocuous game is that it may, particularly when presented in a context, possess some *asymmetries*; and the object is to exploit them. See also pp. 298 ff.

Let us — whether or not we are strongly attracted to the symmetry postulate, and whether or not we are especially attracted to the particular symmetry of the Nash solution — put ourselves in a frame of mind congenial to accepting the "Nash point" as the rational outcome of an explicit bargaining game.[19] Consider now some variants of this game.

[19] The solution proposed by J. F. Nash for bargaining games in which both players have perfect knowledge of their own and each other's utility systems (subjective valuations) is the outcome that maximizes the *product* of the two players' utilities. If all possible outcomes are plotted on a graph whose rectangular coordinates measure the utilities that the two players derive from them, the solution is a unique point on the upper-right boundary of the region. (The point is unique because, if there were two, the two could be joined by a straight line representing available alternative outcomes achievable by mixing, with various odds, the probabilities of the original two outcomes; and points on the line connecting them would yield higher products of the two players' utilities. In other words, the region is presumed convex by reason of the possibility of probability mixtures, and a convex region has a single maximum-utility-product point, or "Nash point.")

A distinguishing feature of this particular "solution" is that it is independent of the exchange rate between the two players' utility scales; it is, in other words, invariant with respect to any fixed weights that we might attach to their respective utilities. And it meets some other conditions, notably including the condition that for any pair of fixed weights (or any exchange rate) relating the two players' utility scales that yields a *symmetrical* region, the upper-right midpoint is the solution; that is, the best point symmetrical as between the two players is the solution. (It is the only solution that does meet all of the specified conditions; Nash showed that any solution meeting his conditions must lead to the outcome that entails the maximum product for the two players' utilities.) For our present purpose we may take this symmetry requirement as the generic characteristic of the solution, and think of the other conditions (axioms) as serving to refine the crude notion of symmetry to the point where a unique solution is guaranteed. See the earlier references (p. 267) to Nash, Harsanyi, and Luce and Raiffa; see also the excellent elucidation of the Nash theory, with criticism, by Robert Bishop, "The Nash Solution of Bilateral Monopoly and Duopoly," to be published. And for an application of the "Nash point" to the theory of arbitration, see Layman E. Allen, "Games Bargaining: A Proposed Application of the Theory of Games to Collective Bargaining," *Yale Law Journal*, 65:660 (April, 1956).

Incidentally, it may deserve to be emphasized that the Nash theory is not just one that does not *need* a means for comparing two players' utility scales — one that, being independent of interpersonal utility comparisons, can get along without them. Rather, since it uses the arbitrariness of the utility exchange rate as a fundamental principle, the theory must be taken to *depend* on the inherent incommensurability of utilities. If the two players' utility scales could in principle be compared, though with difficulty, the Nash theory would not seem an attractive means of obviating difficult comparisons. If in principle utilities were commensurable, there would be little virtue in a theory

First, we are to play the same game in its tacit form. Each of us picks a value along his own axis, and if the resulting point is on or within the boundary, we get the amounts (utilities) denoted by the coordinates we pick. I conjecture that, in the frame of mind I have asked for — a frame of mind that made the Nash point appeal to us in the explicit-bargaining game — we should probably pick the Nash point. Without asking precisely why, let us go on to another variant of the game. This variant is tacit too; but it differs in that we get nothing unless the point whose coordinates we pick is *exactly* on the boundary. We get nothing unless we exhaust the available gains. Caution gets us nowhere; each must choose exactly as the other expects him to. I propose that in our present frame of mind we ought to take the Nash point.

Finally, consider another variant. We are shown the diagram of the game that has just been played and told that we are now to be perfect partners, winning and losing together. Conscious of the fact that our present game is modeled on a bargaining game we are to pick, without communicating, coordinates of a point that lies exactly on the boundary. If we do, we both win prizes — the same prizes no matter what point we succeed in picking together — and if we fail to pick a point on the boundary we get nothing. In this pure coordination game, I conjecture again that we should (would) in our present frame of mind pick the Nash point.

Why? Simply because we need some rationalization that leads to a unique point; and in the context, the bargaining analogy provides it. Unless there is a sharp corner (which is then likely to be the Nash point anyway); or a simple mid-point as when the

---

that relies, in reaching a solution, on the principle of incommensurability. And while the present-day conceptual bases of game theory and of economic theory seem incompatible with interpersonal utility comparisons, the notion of *arbitration* may not be. Economic theory finds it convenient to use a notion of utility that makes utility theory correspond to choice theory, so that one can get "welfare economics" as a free by-product of a theory of economic choice. But if one were to forego this correspondence, for purposes of deriving principles of arbitration, one might be led either to an attempt to measure "utility" in some psychological or physiological way, or to establish legalistically some convention for making a comparison — a convention that, though arbitrary, were compatible with the social purpose of arbitration

boundary is a straight line or circular arc (which again coincides with the Nash point); or some especially suggestive form that seems to point towards a particular point; or unless there is an impurity (such as a dot on the boundary, from a printer's error, or a single point whose coordinates are whole numbers, and so forth), we may be led to search for a "unique" definition of symmetry to fall back on. And Nash-type symmetry is as plausible as any I can think of — not as simple as some (like the intersection with a 45° line from the origin of the diagram and others of that ilk), but less ambiguous on its own level of sophistication.

And, if the Nash point appeals to us powerfully in the bargaining game, it must do so because we are confident that it appeals equally to our partner who in turn we believe to be aware that our views coincide. It must therefore appeal to us in the pure-coordination game as a unique point that the partner will consider to be obviously obvious.

What does this prove or suggest? I am not arguing for the Nash point. I am arguing rather that the appeal of the Nash point to a game theorist (as introspective game player) may be the reverse of the sequence I have just run through. It may be the focal quality of the Nash-point in the pure coordination game — the unequivocal usefulness of a uniquely defined symmetry concept, when no nonmathematical impurities are available to help — that makes it a controlling influence in the tacit and terribly cooperative boundary-line variant of the game; that in turn makes it a reliable guide in the less demanding tacit bounded-area variant of the game; and that in turn takes the heart out of any player in the explicit bargaining game who might hope that expectations could focus anywhere else.

In other words, by postulating the *need for coordination of expectations*, we seem to have a theoretical basis for something like the Nash axioms. What a theory like Nash's needs is the premise that a solution exists; it is the observable phenomenon of tacit coordination that provides empirical evidence that (sometimes) rational expectations can be tacitly focussed on a unique (and perhaps efficient) outcome, and that leads one to suppose that the same may be possible in a game that provides nothing but mathematical properties to work on. The Nash theory is

vindication of this supposition — complete vindication if it dominates all competing mathematical solutions in terms of mathematical esthetics. The resulting focal point is limited to the universe of mathematics, however, which should not be equated with the universe of game theory.

APPENDIX C

RE-INTERPRETATION OF A SOLUTION CONCEPT FOR "NONCOOPERATIVE" GAMES

The pure common-interest game, or coordination game, may add insight into the reasoning behind certain solution concepts in game theory, particularly that of *solution in the strict sense* for the "noncooperative" game. By "reasoning that lies behind these concepts" I mean the reasoning that is imputed to the rational players to whom the concepts should appeal.[1]

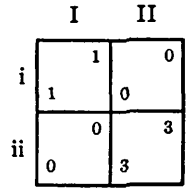

FIG. 25

The tacit games represented in Figs. 25 and 26 are said to have a *solution in the strict sense.* (In Fig. 26 a choice of either second or third strategy for each player constitutes the solution.) The definition of such a solution, given by Luce and Raiffa, is as follows: "A non-cooperative game is said to have a *solution in the strict sense* if: (1) There exists an equilibrium pair among the jointly admissible strategy pairs. (2) All jointly admissible equilibrium pairs are both interchangeable and equivalent." [2]

[1] "Noncooperative" is the traditional name for the game without overt communication. Unfortunately it may suggest that cooperation is absent when communication is absent. As indicated in Chapters 3 and 4, cooperation — reciprocated and taken for granted by each side — is an essential element, even a dominant element, in many tacit nonzero-sum games.

[2] *Games and Decisions*, p. 107f. This particular solution concept is akin to,

	I	II	III
i	1 1	0 0	0 0
ii	0 0	3 3	3 3
iii	0 0	3 3	3 3

FIG. 26

An *equilibrium pair* is a pair of strategies for the two players such that each is the player's best strategy (or as good as any other) that can be coupled with the other's. A *jointly admissible* strategy pair is a pair that is not jointly dominated by another pair; that is, it yields a pair of payoffs that are not both inferior to the payoffs in some other cell. Equilibrium pairs are *equivalent* if, for each player separately, they yield equal payoffs; equilibrium pairs are *interchangeable* if all pairs formed from the corresponding strategies are also equilibrium points. (They are therefore equivalent and interchangeable only if all pairs formed from the corresponding strategies are equivalent.) Thus the strategy pairs (ii, II), (iii, III), (ii, III), and (iii, II) in Fig. 26 denote equivalent, interchangeable, jointly admissible equilibrium pairs.

Luce and Raiffa, immediately after this definition, add the following comment, which can serve as our point of departure: "The second condition prohibits *confusion* in the case of non-unique jointly admissible equilibrium pairs." (My italics.)

It is precisely this problem of *confusion*, or *ambiguousness*, that was at the heart of the coordination game in Chapter 3. The game in Fig. 27 does not have a *solution in the strict sense*. The second and third strategies for the two players are not interchangeable and equivalent — they do not yield equivalent pairs in all four combinations. There is no difference of interest between the two players in their choice of strategies; there is simply cause for confusion. In Fig. 25 they know exactly what

but distinct from, that proposed by J. F. Nash in 1951. For a comparison of several related solution concepts see Chap. 5 of Luce and Raiffa, and J F Nash, "Non-cooperative Games," *Annals of Mathematics*, 54·286–295 (1951)

	I	II	III
i	1 1	0 0	0 0
ii	0 0	3 3	0 0
iii	0 0	0 0	3 3

FIG. 27

strategies to choose; in Fig. 26 they know as well as they need to; in Fig. 27 they do not. Failure to coordinate in Fig. 27 condemns them to zero apiece, and without a clue to coordination they may be supposed to have a fifty-fifty chance of winning 3 apiece, for an expected value of 1.5.

Why is it that (ii, II) is the indicated solution in Fig. 25, rather than (i, I)? An offhand answer is that the payoff is better for (ii, II) than for (i, I). But this is only part of the answer. Another part emerges if we look at Fig. 28, which is like Fig. 25

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

FIG. 28

in preference ordering but different in absolute strengths of preference. In Fig. 28 it looks as though the important thing is not to achieve 10 rather than 9, but 9 or 10 rather than zero. Roughly speaking, the two equilibrium pairs are nearly equivalent but not interchangeable; and though the players may be little concerned about whether they get 9 or 10 they are very much concerned not to get zero. Their main interest is to avoid "confusion."

They need to find some clue, or rule, or instruction to coordinate their choices. In a game as abstract as the matrix in Fig.

28, there is little to guide them but the numbers; and between the alternative rules of picking the lesser pair or the greater, the latter probably has more plausibility. We might ask how much it is worth to the players to have an extra dollar attached to (ii, II) by comparison with (i, I); it is worth a great deal as a signaling device and just a little as extra money. It is the difference between 9 and 10 that makes it possible to coordinate choices. In Fig. 29, if we suppose that they can find no rule

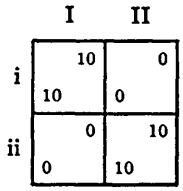

FIG. 29

for coordination, their expected value is presumably 5 apiece.

(Actually the game in Fig. 29 *if presented in the matrix as shown* may not cause difficulty. The empirical results of Chapter 3 imply that it need not. A specific matrix permits left-right, upper-lower, first-last-middle distinctions. For our present purpose, we must suppose that the strategies occur to the players in such form and with such labels that rational players are intellectually incapable of ordering them unambiguously. A completely foolproof or geniusproof clueless game would presumably have to have scrambled labels and a perfectly symmetrical set of payoffs. Incidentally, a tacit game with infinitely many strategies apparently has no "pure" form; an infinity of strategies could only be presented to the players by means of a generating formula, and any generating formula is likely to offer the players some means of ordering the strategies.)

The situation may not be very different if we suppose that the strategy pair (ii, II) is underlined, printed bold face, has arrows pointing toward it, or has a footnote saying that in case of confusion the management suggests a choice of (ii, II). What the players need is *some* signal to coordinate strategies; if they cannot find it in the mathematical configuration of the payoffs,

they can look for it anywhere else. And strategies may occur in such fashion, or with such labels or connotations, as to provide a potential basis for ordering them or sorting them that rational players find useful.[3]

The suggestion of this appendix, then, is that an important property enjoyed by a "solution in the strict sense" — a reason why rational players might select it — is a signaling power, a means of tacit communication, that is available to the two players to facilitate their tacit cooperation when failure to coordinate choices would be serious. This is of course not the only significant property of such a solution; but it may be an important part of the rationale for a player's choosing it.

Another way to make this point is that we could, in games like those presented in this paper, prescribe communication arrangements with certain communication costs and analyse the games to see whether communication is worth the cost and what messages sent over what channels would constitute the "solution." The "clues" under discussion in this paper would then appear to be so much free communication to be taken advantage of; and it is an empirical question what free communication a rational player should be able to find and take for granted. Just as esthetic or syntactic constraints on a language help to eliminate garbles in a badly transmitted message, esthetic or dramaturgical constraints, casuistic or geometric constraints, can help to eliminate ambiguousness in a situation where tacit concerted choice is required.

The point can be pressed further. Consider the game in Fig. 30. Again assume that the strategies occur in a way that makes ordering them intellectually impossible for rational players, specifically, not in the form of a particular square matrix, not

[3] The type of "rationality" or intellectual skill required in these games is something like that required in solving riddles. A riddle is a context in which one is invited to search for a clue, the rules being that the clue must not be too hard to find nor too easy. (One must at least be able to recognize that he should have got it, when it is pointed out to him.) A riddle is essentially a two-person problem; the methodology of solution depends on the fact that another person has planted a message that in his judgment is hard to find but not too hard. In principle one can neither make up nor solve riddles without empirical experience; one cannot deduce *a priori* whether a rational partner can take a hint. "Hint theory" is an inherently empirical part of game theory.

	I	II	III	IV
i	10 10	0	0	0
ii	0	10 10	0	0
iii	0	0	9 9	0
iv	0	0	0	10 10

FIG. 30

labeled with numbers or letters, or — if they are labeled — with the labels scrambled separately for the two players. There it would appear that if no better means of coordinating can be discerned, the "solution" may be the strategy pair (iii, III) with payoffs of 9 apiece. This is the least desired among the equilibrium points, but it enjoys uniqueness while the others offer confusion; it provides a clue to concert choices. In terms of the *payoff* structure alone (that is, without introducing "labels," prefabricated matrices, or any other details outside the pure quantitative structure of the game), it is hard to see that this solution is much less, if at all less, compelling than the one in Fig. 31,

	I	II	III	IV
i	9 9	0	0	0
ii	0	9 9	0	0
iii	0	0	10 10	0
iv	0	0	0	9 9

FIG. 31

although the latter meets the Luce-Raiffa definition and the former contradicts it.[4]

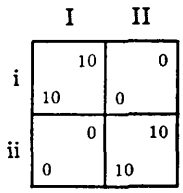

FIG. 32

The games in Figs. 32 and 33, neither of which has a solution in the strict sense, seem to represent the same point. It "looks as though" the players have an argument for choosing (ii, II) in Fig. 33 One argument might be that, in the absence of any way of knowing whether to aim for (i, I) or (ii, II), one should consider what insurance he can fall back on. The row chooser gets nothing if he wrongly chooses the upper row, he gets 5 if he wrongly chooses the lower row, "wrong" meaning that he fails to rendezvous with his partner for 10. He might then choose the lower row arguing that he does so because he will at least get 5 if he does not get 10, and his chances of getting 10 are no worse with this choice. Perhaps this is all that "rationality" requires of him; but it might be more perceptive to reason as follows.

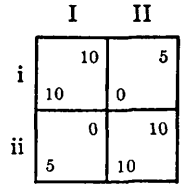

FIG. 33

"Comparing just (i, I) and (ii, II) my partner and I have no way of concerting our choices. There must be some way, however, so let's look for it. The only other place to look is in the cells (ii, I) and (i, II). Do they give us the hint we need

[4] Empirical evidence for these and similar games can readily be obtained for himself by any reader who wants to pursue the point.

to concert on 10 apiece? Yes, they do; they seem to "point toward" (ii, II). They provide either a reason or an excuse for believing or pretending that (ii, II) is better than (i, I); since we need an excuse, if not a reason, for pretending, if not believing, that one of the equilibrium pairs is better, or more distinguished, or more prominent, or more eligible, than the other, and since I find no competing rule or instruction to follow or clue to pursue, we may as well agree to use this rule to reach a meeting of minds."

In this case the players are not choosing their second strategies because 5 is preferable to 0. They have no serious expectation of getting 5. They are *using* the configuration of fives and zeros as a *clue* to coordinating actions. It is *useful* to the players — and each recognizes that the other recognizes that it is useful — to take note of where the fives are, but only as a step in the process of coordinating intentions. The tendency for the matrix in Fig. 33 to "converge" on (ii, II) is in principle the same as if the printed matrix had arrows pointing toward the lower-right corner, arrows with no logical role or authority other than the power of suggestion and hence the ability to coordinate expectations.[5]

CONFLICTING INTEREST

We can consider now the case of coordination mixed with conflict. Figures 34 and 35 portray games that have equilibrium points, two of them both jointly admissible, without a "solution in the strict sense" because the equilibrium pairs are neither equivalent nor interchangeable.

The coordination problem in the first of the two is apparently "insoluble" in its purely abstract form, that is, without labels on the strategies; there appears to be at best a random chance

[5] Assuming that a player does choose ii or II, it may be worthwhile to find an operational way of discriminating between motives for choosing it, even if only to make sure that the concept is operational. As between the two motives mentioned — the "insurance" motive and the "coordination-clue" motive — we might distinguish as follows. We offer a player alternative games like Fig. 33 that differ only in substituting values ranging from 0 to 9 for the 5's in that matrix, leaving the 10's and zeros as they are. We then ask him to "value" the games for us — to indicate how much he would pay for the opportunity to play the game with a live partner and real money payoffs. (Alternatively

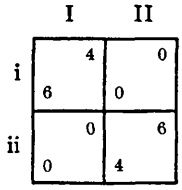

FIG. 34

of achieving either of the jointly admissible (efficient) outcomes.[6] The second may not be insoluble. Each player would rather accept his "second-best" equilibrium point than fail to coordinate at all; they have a common interest in cooperating to find a clue to common choice. Why not take the clue contained in the other cells, which seems to point toward (ii, II)?[7]

---

we ask him how much he'd pay for the privilege of playing the different variants in place of the one with 5.) If his response is fairly insensitive to variations in that particular payoff as long as it is positive, and if nevertheless he attaches a high value to the game with some positive payoff and attaches something like a random-strategy expected value for the game with zeros as in Fig. 32, we can conclude that the lower-left and upper-right payoffs are mainly of interest to him as signals. If, for example, he bids \$9.50 for a chance to play the game in Fig. 33 (implying, perhaps, a 90 percent expectation that Column will choose II), \$8.65 for the game with 5 replaced by 1 (implying an 85 percent expectation of II), and \$9.95 for the game with 5 replaced by 9 (implying a 95 percent expectation of II), and, finally, \$5 for the game as in Fig. 32 (implying a random expectation as between I and II), we could conclude that the function, or value to the player, of the upper-right and lower-left payoffs is largely that of coordinating clue. If instead he bids amounts that imply probabilities between I and II that are invariant, or nearly so, with respect to the upper-right and lower-left payoffs, and particularly if he bids the arithmetic mean, the insurance interpretation would be indicated. (Note that the adjectives "upper-right" and "lower-left" are only author's shorthand here; they have no meaning to the player since we are considering the case of *unlabeled* strategies, which must not be presented in a square matrix, or with labels like "i" and "ii" — or, if they are, must have been labeled by a random process separate from the random process that allocated labels or positions for the other player. Specifically, Row must not know whether Column's matrix looks like Fig. 33 or instead has the columns interchanged with the low-value payoffs in upper-left and lower-right.)

[6] See the footnote on p. 286 for a discussion of a similar matrix when the premise of pure abstration is relaxed.

[7] The game in Fig. 35 does have another equilibrium point, consisting of an 80:20 mixed strategy for Row and a 40:60 mixture for Column. It yields them payoffs of 3.6 apiece, and is therefore jointly dominated by the upper-left and lower-right cells.

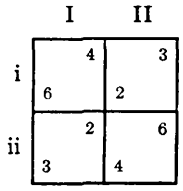

FIG. 35

For one of the players this is not the most advantageous outcome, but beggars cannot be choosers when fortune gives the signals. What other clue is there? It might be equally *fair* to use the negative of this clue; just as it would be equally fair, if arrows pointed toward (ii, II) and away from (i, I), to treat the feathers as the signal rather than the arrowheads. But fairness cannot help; in fact it makes coordination impossible. If all clues are equally plausible in reverse, we are back to confusion. Only a *discriminatory* clue can point to a concerted choice, denying the discrimination is denying the premise that a clue can be found and acted on jointly to achieve an efficient outcome in the face of conflicting preferences.[8]

Here again the most potent clues may be those that we admit when we go beyond the mathematics of the payoff matrix. If we are driving toward the same intersection on perpendicular roads on a desert where no legal system determines right-of-way, and dislike and distrust each other and recognize that there is no moral obligation between us, the one approaching on the other's left may nevertheless still slow down to let the other through first, to avoid emergency stops at the intersection; and the other driver may anticipate this.[9] The conventional priority system lacks legal or moral force; but it is so expedient when coordination is needed that the one discriminated against may yield to its

[8] The power of similar mutually perceived signals seems to lie behind the concept of "psychological dominance" used by Luce and Raiffa to discuss the appeal in certain games of a jointly inadmissible equilibrium point. See *Games and Decisions*, pp. 109–10. See also the footnote on p. 286 for a comment on a similar game.

[9] A conflict-of-interest problem of this type — two cars approaching an asymmetrical narrow place in the road from opposite directions — was included in the questionnaire described in Chapter 3. The results bore out the general principle, but were omitted for brevity from Chapter 3.

discipline, recognizing that he should be grateful for an arbiter, even though it discriminates against him, and recognizing also that he is trapped by the other's acceptance of the signal and expectation that both will comply. By this reasoning, as developed in Chapter 3, the game in Fig. 34 may be soluble when presented in a *particular* matrix form to both players (that is, presented just as shown in Fig. 34), or when the winning strategy pairs are *labeled* "heads" and "tails," i, ii, I, and II, and so forth.

MANIPULATION BY A THIRD PARTY

Incidentally, all of these games requiring coordination, both those with conflicting preferences and those with preferences that coincide, might be substantially subject to the control or influence of a mediator. If we give a third player power to send messages to the original two tacit players, he is in a good position to help them; he is even in a good position to help himself if he gets a payoff that depends on the pair of strategies that the original two players choose. A benevolent mediator makes the pure common-interest game trivially easy; a mediator has an arbitrary power of justice in a game like that of Fig. 34; [10] a mediator is in a strong "third player" position in the game in Fig. 36, where the entry in parentheses is the payoff to the

	I	II	III	IV
i	5 (2) 6	0	0	0
ii	0	7 (3) 9	0	0
iii	0	0	6 (4) 5	0
iv	0	0	0	9 (1) 7

FIG. 36

[10] Recall problem no. 8 on p. 62 of Chapter 3, involving lost and found money and a self-appointed mediator.

mediator (or communication monopolist) who is in a position to give instructions — suggestive only, not authoritative — to the other two players.

INTERPRETATION OF THE PAYOFFS

As a final point it may be noted that, for the line of reasoning developed here, it does not matter whether we interpret the payoffs as objectively measurable entities, such as money or homogeneous goods, or as "utilities" in the sense now familiar in game theory. It does not depend on each person's knowledge of the strengths of the other's preferences, as long as the nominal payoffs are known. (If both the objective values and the utility values were known, and were not proportionate to each other, the "signals" might lose some force; the problem of confusion or ambiguousness would be aggravated.)

NUMBER OF PLAYERS

The discussion here has considered only two-person games, except for brief consideration of a third player who may be in a nontacit role. But the problem can be extended to any number of players, with the rewards depending either on unanimous choice or on some kind of majority or plurality choice or successful coalitions (somewhat analogous to the lines of the actual questionnaire procedure described in Chapter 3). The problem of ambiguousness may then become more serious, and the coordination aspect of the game may become even more relevant to the rationale of a "solution." It is probably in the realm of more-than-two-person games that coordination theory is most relevant of all, games involving the formulation of coalitions. Study of the signals and communication channels in coalition formation appears to be a fruitful meeting ground for game theory and sociology.

CONCLUSION

In summary, coordination-game theory suggests that the "solution in the strict sense" of a tacit nonzero-sum game is to be

understood partly, and in some cases largely, by reference to its signaling qualities. Since other sources of signals may be present *even in the purely mathematical formulation of the game,* the particular qualities of the "solution in the strict sense" are but one of many potential determinants of a "rational solution." It is partly an empirical question, not solely a matter of deduction a priori, what signals can be appreciated.