The Theory of Games and the Problem of International Cooperation

R. HARRISON WAGNER
The University of Texas at Austin

The Prisoner's Dilemma game, Rousseau's image of the Stag Hunt, and the concept of a security dilemma have all been used to support the argument that much international conflict is the result of anarchy at the global level rather than the aggressive intentions of governments. This article argues that the Prisoner's Dilemma and the Stag Hunt are usually inadequate models of the problem they have been used to illuminate, and that a security dilemma as commonly defined need not have the implications that are ascribed to it. It also argues that developing more adequate models of the general problem of enforcing agreements in a condition of anarchy will help us to understand better why international cooperation is more easily achieved in some areas than in others.

Because of the absence of a global government, it is frequently argued, international cooperation is difficult to achieve, and conflicts often occur not because of any state's aggressive intentions, but because in a condition of anarchy, even states with purely defensive motives will find it difficult to avoid them. There are a variety of models, metaphors, and strands of argument designed to support such a conclusion. One is Rousseau's image of the Stag Hunt—cooperation is everyone's first choice, but the consequences of cooperating when others do not are disastrous (Jervis, 1978). Uncertainty about others' intentions therefore leads to defensive noncooperation. Another is the concept of the security dilemma—actions taken by one state to increase its security diminish the security of others, thereby leading to conflict. A third is the familiar Prisoner's Dilemma—everyone has a dominant strategy to defect, but the result is an outcome worse for everyone than if everyone had cooperated (Jervis, 1978; Snyder, 1971).

I will argue that the Stag Hunt and the Prisoner's Dilemma are often inadequate models of the problem they have been used to illuminate, and that a security dilemma as commonly defined need not have the implications that are often ascribed to it. I will also argue that developing more accurate models of the general problem of enforcing agreements in a state of anarchy will help us to understand better why international cooperation is more easily achieved in some areas than in others.

I would like to thank the following people for reading and commenting on an earlier version of this article: Jay Budziszewski, William Galston, Jack Levy, and Cliff Morgan.

Prisoner's Dilemma Models

It will be useful to begin the discussion by considering the Prisoner's Dilemma game. For two reasons, this game plays a prominent role in discussions of the problem of cooperation in the absence of an enforcer of agreements. First, it illustrates the general point that equilibrium outcomes in noncooperative games can be suboptimal. In doing so it makes us think about whether only equilibrium outcomes can be solutions to noncooperative games. Thus it focuses our attention on the problem of what sort of outcomes rational players will arrive at in any noncooperative game.

Second, the preference orderings in the Prisoner's Dilemma game seem to represent the preferences of individuals (or governments) in many situations in which it is possible for some to refuse to cooperate while others are willing to cooperate. For example, it is plausible to identify the "C" and "D" choices of the Prisoner's Dilemma game with "no tariff" and "tariff," or "do not arm" (or "disarm") and "arm." It is sometimes plausible to identify them with "do not attack" and "attack." Indeed, it is illuminating to consider the set of international boundaries as the result of an international agreement, which implies a similarity of form between the problem of war and the general problem of international cooperation.

Of course, "cooperate" and "defect" may be plausible labels for governments' choices, and yet their preferences over the four possible combinations of choices may not be those of the two prisoners. They may, for example, correspond to those of Rousseau's Stag Hunt.[1] However, there

[1]A Prisoner's Dilemma game is represented in Figure 1 below. Jervis's (1978) game theoretic representation of

are other ways in which the Prisoner's Dilemma may fail to model accurately the situations to which it has been applied, even when governments' preferences do correspond to those of the prisoners. An examination of these other inaccuracies will demonstrate problems with not only the Prisoner's Dilemma game, but also the entire class of 2 × 2 games as tools of analysis.

Another important way in which the classic Prisoner's Dilemma game may fail to represent reality accurately is that the situation it models may be a recurring one. In an article that stimulated some important research, Shubik (1970) discussed a Prisoner's Dilemma game that is repeated an unknown number of times, which gives rise to a new game in which mutual cooperation can be an equilibrium (Axelrod, 1981; Taylor, 1976). However, Shubik (1970, p. 190) also said of this iterated Prisoner's Dilemma game, "I claim that for most problems of interest the model is still not rich enough to capture a useful abstraction of human affairs."

This cautionary statement has not prevented some scholars who have been influenced by Shubik's article from suggesting that the iterated Prisoner's Dilemma game is a general model of the problem of achieving cooperation in the absence of an enforcer of agreements. For example, in summarizing a recent article on this game, Axelrod (1981) says, "This article investigates the conditions under which cooperation will emerge in a world of egoists without central authority." The investigation of under what circumstances rational players will cooperate with repeated plays of the Prisoner's Dilemma game is an important line of research with possible applications to the problem of international cooperation. However, I will argue that not even the Prisoner's Dilemma supergame is an accurate model of many of the situations to which the Prisoner's Dilemma has been applied, and this for reasons more fundamental than the ones given by Shubik in his original article on the subject.

Let us consider, then, other possible ways in which the Prisoner's Dilemma may be an inaccurate model of a situation. In addition to the assumptions just mentioned, there are three other conditions that must be met for a situation to be represented accurately by the Prisoner's Dilemma game: 1) There must be only two actors; 2) each must have one and only one opportunity to choose between the alternatives, C and D, before payoffs are received; and 3) each must choose in ignorance of the choice made by the other.[2] Clearly the last condition often is not met; governments generally know that other governments have or have not violated agreements when deciding whether to continue to cooperate. What happens if we relax that condition?

Having the Last Word

The matrix in Figure 1 is an example of a standard Prisoner's Dilemma game. Figure 2 contains the same game in extensive form, without the stipulation that the players choose in ignorance of the other's choice. (To represent that condition in Figure 2, we would merely draw a line encircling player 2's two choice nodes, indicating that he could not distinguish between them.) Now, of course, whoever moves first still has only two strategies, but the player who moves second has four. This is no longer a standard Prisoner's Dilemma game, but it is still a game with a dilemma. Player 1 does not have a dominant strategy, but player 2 still does. Player 1 must therefore anticipate that player 2 will defect no matter what player 1 does, and therefore player 1 will choose to defect first. Thus it would be mere pedantry (although technically correct), to say that this situation is inaccurately modelled by the standard Prisoner's Dilemma.

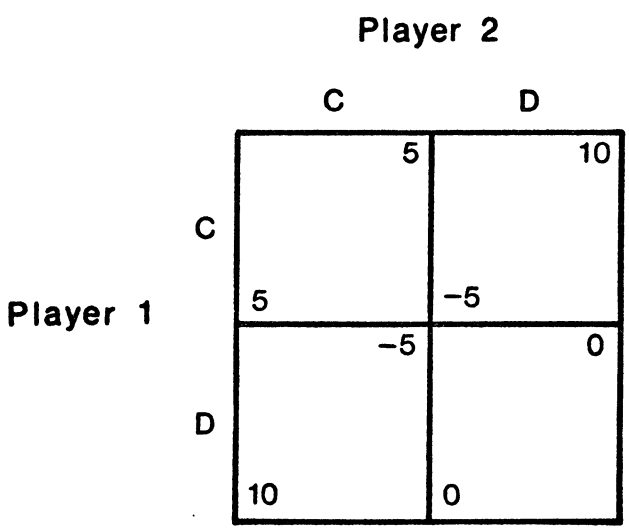
However, this game demonstrates what is genuinely crucial for the existence of such a dilemma, and that is the second condition stated above: that each player can choose only once. This fact makes it possible for a C choice to be followed by a D choice, which is the essence of the problem. Thus the crucial feature in the well-known story of the two prisoners is not that they were separated, but that once one had confessed while the other kept silent for a time, the latter was deprived of another opportunity to confess. After all, when he found out how he was being charged, he would find out also that the other had confessed. Thus the important fact was that there was a deadline after which his own confession could no longer help him.

Before exploring the significance of this fact, let us draw two further lessons from this game. First, would it make sense to try to overcome the dilemma the game embodies by employing Howard's (1971) notion of a metagame? To do so we would first construct its normal form and then imagine

Rousseau's Stag Hunt merely reverses the order of preference between the CC and DD outcomes for both players.

[2]Snyder states that the assumption is that players move simultaneously, which is not correct. He also assumes that with sequential moves one can continue to use the 2 × 2 matrix as a tool of analysis, which is also incorrect (Snyder, 1971, p. 69; Snyder and Diesing, 1977, pp. 44, 164).

Figure 1



that first one and then the other player could select his strategy after knowing what strategy had been selected by the other, thereby generating a whole series of normal forms of games based on this one—Howard's metagames. But the brute fact is that in *this* game, player 2 *already* moves after player 1. How could any metagame alter the fact that once player 1 has chosen, player 2 knows his choice and is free to pursue his own interests without worrying about player 1's choice at all? Once player 1 has chosen, player 2 *knows all he needs to know* to choose.

This reasoning suggests that the notion of a metagame is inappropriate here. But this game is identical to the majorant (for player 2) of the Prisoner's Dilemma. From this one should conclude that Howard is mistaken to equate von Neumann and Morgenstern's (1944, pp. 100-101) majorant game with a metagame, and that the central problem with the notion of a metagame is that it fails to take the extensive form of the game as its most basic representation.

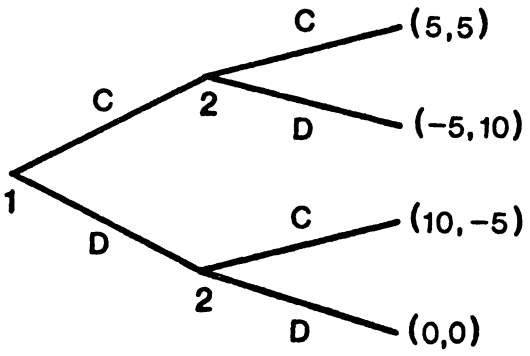
Second, this game (considered now as the majorant of the Prisoner's Dilemma) also demonstrates why people will sometimes cooperate even in single plays of the standard Prisoner's Dilemma under laboratory conditions. For player 2 to choose D after player 1 had chosen C violates many people's ethical standards. Thus even if one succeeds in inducing preferences over the various payoffs of the game that correspond to the Prisoner's Dilemma orderings, players' ethical norms may lead to a different preference ordering when these payoffs are anticipated as the result of taking advantage of the other player (Sen, 1977). Actual plays of a standard Prisoner's Dilemma under laboratory conditions may therefore really be Bayesian games (Harsanyi, 1967-68), in which the players must try to estimate the probability that each will play ethically and that each believes the other will.

But let us return to the main subject and consider the significance of the fact that in both versions of the Prisoner's Dilemma discussed so far, each player has only one opportunity to choose. What happens if we relax that assumption? Let us extend the game tree in Figure 2 to the right; does that make any difference? It clearly does not if the last portion of it contains only segments like the one represented in Figure 2. For then, on the last move the last player would defect and therefore on the next-to-last move the other player would as well, and so on back to the beginning.

But why must the end of the tree look like that? Suppose, for example, that the two celebrated prisoners each had a friend in the District Attorney's office who would inform him whether the other had confessed. Then any confession would be immediately matched by the other's confession. The game tree would then look like the one in Figure 3. Its normal form is given in Figure 4. Now there are two equilibria (strategy combinations 2,2 and 3,4), but one is Pareto superior to the other. It is the result of both players' following the conditional strategy, "cooperate, then defect if the other defects."

Thus in any situation in which the alternatives are the classic C and D choices of the Prisoner's Dilemma, and in which players choose with full knowledge of each other's choices, D will be an

Figure 2

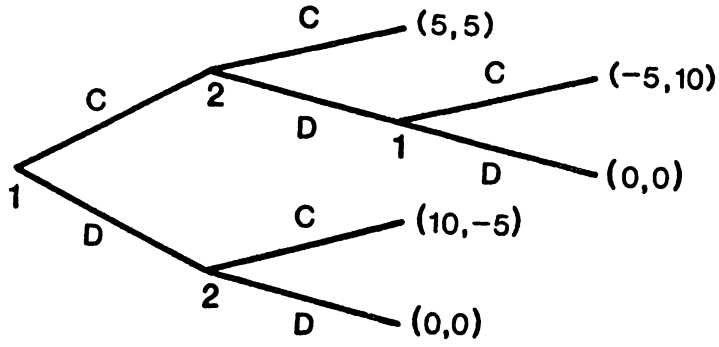


optimal choice for player 1 only if he believes that his choice will be either the last or the next to last choice that can be made by either player. Stated more positively, in any situation in which players choose one after the other, with full knowledge of each other's choices, conditional cooperation will be the optimal strategy as long as no one can count on having the last choice.

The similarity between this conclusion and the analysis of the indefinitely iterated Prisoner's Dilemma should be obvious. However, it is important that they not be confused. For the choices represented in the game tree in Figure 3 are not repeated plays of a game; they are repeated choices in a single game. In the iterated Prisoner's Dilemma it is difficult to determine which strategies are optimal (Axelrod, 1981). In this game it is obvious that conditional cooperation is a Pareto optimal equilibrium. If I offer a check to a perfect stranger as payment for his delivery of some heroin for me, that is a single play of a classic Prisoner's Dilemma game: he may not deliver the heroin, my check may be no good, and neither of us can rely on the state to enforce our agreement. If I regularly employ him for that purpose, that is a Prisoner's Dilemma supergame. But the two prisoners with friends in the District Attorney's office are not playing either of those games.

To illustrate both the difference and its importance, suppose we assume that the classic Prisoner's Dilemma game is an accurate model of an arms control agreement (Brams et al., 1979; Jervis, 1978; Snyder, 1971). The plausibility of any attempt to represent it as such is based on the advantage that would go to a country if it could arm itself while the other remained disarmed. One possible outcome of a single play of this game, therefore, would be that one country achieved whatever gains would accrue to it from such military superiority. A supergame corresponding to this game might therefore lead to a sequence of

Figure 3



STRATEGIES

Player 1:

1. Cooperate whatever 2 does
2. Cooperate and then defect if 2 defects
3. Defect

Player 2:

1. Cooperate whatever 1 does
2. Cooperate if 1 cooperates, defect if 1 defects
3. Defect if 1 cooperates, cooperate if 1 defects
4. Defect whatever 1 does

such occurrences. And a tit-for-tat strategy in such a supergame would be to respond to the cheating of one's opponent in the current agreement by cheating in the *next* arms control agreement (whose payoffs are assumed to be the same), and cooperating in the *following* arms control agreement only if the other side allowed itself to be exploited.

Moreover, it is essential that the players anticipate an indefinite series of such arms control agreements; otherwise they will want to cheat in the last one, and so on back to the beginning. Thus Brams et al. (1979), ignoring the problem of what iteration of the game means, imagine that an arms control game is played twice, with each country cooperating conditionally and each able to determine only with a certain probability whether the other cheated in the first round. They are then astonished to discover that each government would want to have very poor intelligence-gathering abilities—both would have an interest in being unable to be sure that the other had in fact not cheated. But this situation occurs entirely because conditional cooperation does not lead to an equilibrium outcome when the game is repeated only once. Since Brams et al. have denied their players the opportunity simply to cheat (by assuming that they will cooperate conditionally), the players are reduced to having to try to avoid noticing that the other side cooperated.

The Prisoner's Dilemma supergame is plainly inapplicable to the problem of arms control, but the game in Figure 3 may be relevant. The difference is that in the latter game, one has both an immediate opportunity and an obvious incentive to respond to one's opponent's cheating by rearming oneself. If one can detect an opponent's cheating and has time to rearm before the opponent can exploit that advantage, an arms control agreement can be stable. This, of course, will be a revelation only to those who thought that the Prisoner's Dilemma was a general model of the problem of cooperation in a condition of anarchy.

However, this rather obvious conclusion may have a not-so-obvious implication. In spite of the absence of a central political authority, there is an extensive amount of organized cooperation among sovereign states (Young, 1980). An important question, therefore, is why cooperation is easier to achieve in some areas than others. Perhaps the ability of states to respond more quickly and effectively to violations of international monetary or trade agreements, for example, than to violations of arms control agreements helps to account for the greater prevalence of the former. It is on this basis that Nicholson (1972)

Figure 4

Player 1 \ Player 2	1	2	3	4
1	5 5	5 5	10 −5	10 −5
2	5 5	5 5	0 0	0 0
3	−5 10	0 0	−5 10	0 0

explains the tendency of oligopolists to compete in advertising rather than price. The point is undoubtedly of much more general significance.[3]

Intersubjective Knowledge of Utilities

We have seen that Prisoner's Dilemma models, by assuming away the possibility of conditional strategies, ignore the possible existence of cooperative equilibria even when governments' preferences over outcomes are those of the Prisoner's Dilemma. However, one might object that this conclusion assumes intersubjective knowledge of utilities, which normally does not exist, and thus exaggerates the prospects for achieving mutually optimal outcomes in the absence of government (Jervis, 1978, pp. 167-169). Perhaps with incomplete information about utilities, the possibility of cooperative equilibria becomes sufficiently remote that the Prisoner's Dilemma can at least be taken as an illuminating metaphor, if not a fully accurate model of the basic problem of international politics.

Such an objection would be incorrect. With intersubjective knowledge of preferences, the distinction between an equilibrium in a game of perfect information (i.e., a game in which players choose sequentially with full knowledge of the choices made by others) and a game without perfect information is not important.[4] However, when players are uncertain of each other's preferences, the distinction is very important. In a game in which each player must choose without knowing the other's choice (as in any 2 × 2 game), the only information they have that can lead them to mutually optimal choices is information about each other's preferences. If that information is degraded, the likelihood that they will be able to choose optimally is diminished. But in games in which the players have information about each other's choices, any uncertainty about each other's utilities can be much less important. Thus even Jervis's Stag Hunt can be a misleading model of the problem of international cooperation, since it may greatly exaggerate the difficulty of coordinating expectations when there is uncertainty about utilities (Jervis, 1978, pp. 167-169).

[3]Nicholson's analysis is based on an extended and continuous version of the game represented in Figure 3. Although he speaks of the Prisoner's Dilemma game, and Taylor (1976, p. 96) seems to say that Nicholson is analyzing a Prisoner's Dilemma supergame, I think it is clear that Nicholson's model is not a Prisoner's Dilemma supergame because the payoffs are enjoyed continuously.

[4]The term "perfect information" in game theory refers to the knowledge players have of prior moves and not their information about each other's preferences.

The importance of this point can be readily seen by looking at Figure 3. Player 1 may be uncertain about whether player 2 really prefers CC to DD. Even so, he risks nothing by choosing C, since if player 2 then defects, the outcome will be no worse for player 1 than if he had chosen D. But it makes no difference which player is labeled 1, and therefore the same thing can be said about both of them.

However, if we alter the game in Figure 3 slightly, uncertainty about utilities becomes important, and doing so will greatly clarify the meaning of a "security dilemma," and its relation to the Prisoner's Dilemma.

Deterrence and the Security Dilemma

Once we drop the assumption that players choose only once, it becomes possible that not every pair of choices will always lead to the same payoffs. Let us consider again Figure 3 and suppose that player 1's initial choice is between accepting or not accepting an agreement proposed by player 2 that would alter the status quo in some way. It is possible that the consequences of having to respond to player 2's violating that agreement would be worse that the consequences of not signing the agreement in the first place. In that case, player 1's payoff after the second pair of D choices in Figure 3, although greater than -5, might nonetheless be worse than 0 (his payoff after the first pair of D choices). Mutual cooperation would still be an equilibrium outcome; in game theoretic terms, there is no "dilemma." However, achievement of this equilibrium now depends upon intersubjective knowledge of utilities. Player 1 must be able to count on 2's anticipating that if he defects after player 1 had accepted the agreement, the consequence will be a payoff of 0 to player 2, rather than 10; therefore player 2 will prefer to cooperate as well. Otherwise player 1 will prefer to defect on his first move.

With uncertainty about utilities, there are two reasons why player 1 might not be confident about 2's choice. First, player 1 may not be certain that 2 really prefers mutual cooperation to the consequences of defection if he retaliates. Second, player 1 may not be certain that 2 believes that 1 will retaliate, i.e., that 1 prefers to defect rather than cooperate after 2 has defected. Obviously, if player 1 is sufficiently uncertain about either or both these factors, he will choose not to cooperate.

In altering the payoffs in this way, we have already changed one of the most basic assumptions of all Prisoner's Dilemma models—that there are only four distinct outcomes. However, we have come closer to modelling at least some of

the situations to which it has been applied. It is now time to reconsider another basic assumption: that actors have only two alternatives (C and D) between which to choose. If we drop that assumption, we can consider the consequences of a variety of C choices as well as a variety of D choices, each with distinct payoffs attached.

There are many possible implications of such an increased range of choices for the substantive problems to which analysts have tried to apply the Prisoner's Dilemma game. One implication pertinent to the situation just discussed is that another form of cooperation may be possible (in addition to whatever cooperation results from CC): actors can cooperate in increasing each other's incentives to cooperate by arranging for retaliatory moves they can have mutual confidence in. This, of course, is the point not only of inspection systems in arms control agreements, but also exchanges of hostages, the implementation of agreements in small steps, and other such devices. The point of all of them is to arrange either for a situation as represented in Figure 3, or for one with a more precarious equilibrium in which there is nonetheless intersubjective knowledge of utilities.

Choosing the Best Retaliatory Response

In addition, of course, players will act unilaterally to develop retaliatory choices that increase their confidence in each other's continued cooperation. Player 1, for example, will try to develop retaliatory options that player 2 believes player 1 has an incentive to choose, and that clearly lead to worse outcomes for player 2 than continued cooperation. Thus we are now able to take into account that in the real world actors have both an opportunity and an incentive to respond to noncooperation in ways that are not captured in Figure 3, or by the Prisoner's Dilemma supergame. In the heroin example, that one person regularly employs another to deliver heroin is unlikely to be sufficient incentive to prevent cheating, unless the quantities delivered are small or the wages are high. It is much more likely that the employer will find it useful simply to threaten to punish his employee for any cheating. That the transaction is repeated many times is important in justifying the cost of the punishment to the employer, but the iteration of the threat does not constitute a Prisoner's Dilemma supergame.

How best to respond to noncooperation is no more and no less than the problem of deterrence; this can be clearly seen if we give the game represented in Figure 3 a slightly different empirical interpretation. Suppose a set of international boundaries has been established by agreement between two countries, represented by players 1 and 2. Once the agreement has been implemented, enforcement is a continuing problem. At each time period, each government has a choice between continuing to accept the agreement, or overturning it by force. Thus, in Figure 3, player 1's initial choice is not whether to accept a proposed agreement, but whether to continue to abide by it or to overturn it, and player 2 faces a similar choice at his move should player 1 decide on continued cooperation.

Player 1's problem is thus to find a retaliatory response to player 2's defection such that player 2 expects that player 1 will choose it, and to which player 2 clearly prefers continued cooperation. But this is simply the problem of deterring an attack by player 2. It obviously has two components: the credibility of player 1's threat, and the relative severity of player 1's threat (as compared to player 2's evaluation of continued respect for the existing boundaries).

This is admittedly a slightly extended use of the term *deterrence* compared to some common usage, which distinguishes between punishing an attacker (deterrence) and defeating him (defense). Nonetheless, it is clear that one can also deter an attack by having a reliable defense against it, so my usage of the term is not contrary to common sense. The distinction between deterrence by punishment and deterrence by defense concerns the relation between player 1's and player 2's payoffs after the second pair of D choices in Figure 3. To deter player 2 by denying him the fruits of victory is simultaneously to guarantee player 1 a fairly high payoff. Deterrence by punishment severs the connection between the two players' payoffs in case of war, making possible much more severe costs to player 2, but also severe costs to 1 which increase the problem of credibility.

Player 1, then, must analyze carefully the reduced game that consists of the branches of the tree in Figure 3 that follow his initial C choice, with the payoffs as variables. This is the deterrence game, and it should be obvious from Figure 3 that it is absurd to represent it (as is often done) by the 2 × 2 matrix known as "chicken." However, my purpose here is not to analyze this narrow deterrence problem, but to find a way of representing the security dilemma (the existence of which, as is well known, can make narrowly conceived deterrence strategies backfire).[5]

[5]The deterrence subgame tree can be extended further to the right, allowing for second, third, or more "strikes," and other complications. For a full discussion, see Wagner (1982).

The Security Dilemma

To understand what the notion of a security dilemma refers to, let us continue to assume that the problem faced by our two players is whether to continue to abide by an agreement establishing a set of international boundaries or to overturn it by force. In order to simplify the discussion, it will be convenient to simplify Figure 3 slightly (although we should not lose sight of what the complete tree looks like). Let us ignore for the moment the problem of credibility of each player's retaliatory threats and merely focus on a single set of payoffs to both players as the expected consequence of either's defection. We will in any case want to consider variations in these payoffs, which are uncertain due to incomplete information. We should merely not forget they are in each case the result of two factors: the severity of the threat by the retaliator and its credibility, and that changes in them can be the result of choices that affect either factor.

If we snip off the retaliatory C choices, then, and make the payoffs variable, we get the game tree represented in Figure 5. Its normal form is represented in Figure 6. There are three distinct outcomes: continued cooperation (respect for boundaries), labelled *c;* a defection (attack) by 2 followed by defection (retaliation) by 1, labelled *b;* and the reverse possibility, labelled *a.* I will refer to the payoffs to players 1 and 2 from outcome *a* as a_1 and a_2 and similarly for *b* and *c.* Remember also that these are *recurring* choices made in the context of an *existing* set of boundaries. Thus the game tree in Figure 5 can be extended indefinitely in both directions by giving player 1 another choice if 2 elects continued cooperation, and so forth. It will be convenient to examine the game from the point of view of player 2, who is making a choice *before* player 1's choice, which initiates the tree in Figure 5. If player 2 chooses to defect, he gets a payoff identical to b_2. If he chooses to cooperate, the consequences depend upon player 1's choice represented in Figure 5. Thus, whether player 2 chooses to defect or cooperate depends upon his anticipation of the outcome of the game represented in Figure 5, and that is what we will examine, keeping in mind that the game extends indefinitely to the right.

If the game does not end should both cooperate, one might ask to what the payoffs at outcome

Figure 5

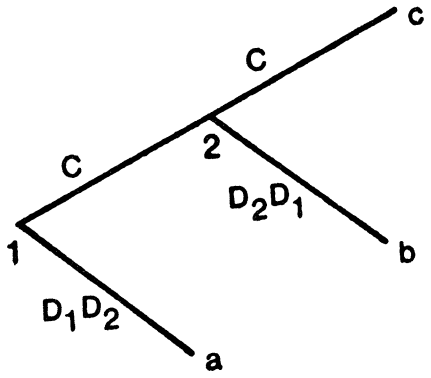


Figure 6

Player 1 \ Player 2	C	D_2D_1
C	c_1 c_2	b_1 b_2
D_1D_2	a_1 a_2	a_1 a_2

c refer. Clearly the benefits of cooperation are enjoyed continuously, and thus the payoffs at c must be regarded as the present value of such benefits as they extend indefinitely into the future (taking into account possible future defections that might result from changes in the game's payoff structure over time).

Let us suppose that $c_2 > b_2$, and $a_2 = b_2$. At his move before 1's opening move in Figure 5, therefore, player 2 will choose to cooperate. Assume also that 2 is confident that $a_1 = b_1$. In that case player 1 will choose C if and only if $c_1 > a_1$. If player 2 is uncertain whether this inequality holds, continued cooperation (though clearly preferable for him) entails a risk—the risk that player 1 no longer prefers to honor 2's control over the territory assigned to him. (Since at one point both players accepted an agreement assigning player 2 that territory, and 1 has not yet moved to overturn it, this uncertainty must be the result of some change that has occurred, perhaps in the military balance or in the decision-making process of 1.) If player 2 can decide more than simply whether to cooperate or defect, but can also develop and communicate to player 1 a variety of retaliatory choices that influence either or both a_1 and a_2, then player 2 confronts an additional decision problem: which such retaliatory alternative would be optimal, in light of its impact on the consequences of 2's choice to continue to cooperate?

More formally, with the assumptions just stated the expected value to player 2 of continued cooperation is:

$$S_2 = p_{21}a_2 + (1 - p_{21})c_2 \qquad (1)$$

where p_{21} is player 2's subjectively estimated probability that 1 will defect, in this case equal simply to the probability that $c_1 < a_1$. The word "security" is widely used in the international relations literature, but is rarely defined. It is plausible to say that S_2 represents player 2's security. Player 2 obviously wants to select and to communicate to player 1 retaliatory moves that maximize S_2. Player 1 faces a similar problem.[6]

Alternatively, it is possible to identify the common notion of "security" with the game theoretic concept of "security level," that is, the worst possible consequence of a strategy choice. Player 2's security level in this game is a_2. If maximizing security in this case is equated with maximization of one's security level, then player 2 will seek merely to maximize a_2. Clearly the effect of doing so is to ignore the deterrent consequences of one's choices, that is, their effect on p_{21}. It is important to keep these alternative security objectives in mind during the subsequent discussion, since they have significantly different consequences for the severity of security dilemmas.

Let us begin by considering a situation in which improvements in one country's defensive position worsen the payoffs to the attacker, but there is no advantage to attacking first. That would be true if:

$$c_2 > b_2,\ c_1 > b_1$$
$$a_1 = b_1,\ \frac{da_1}{db_1} = 1 \quad a_2 = b_2,\ \frac{db_2}{da_2} = 1 \qquad (2)$$
$$\frac{da_1}{da_2} < 0 \quad \frac{db_2}{db_1} < 0$$

In that case, maximizing a_2 is equivalent to minimizing a_1, which also minimizes the probability that $c_1 < a_1$, and hence p_{21}. Thus in this case maximizing player 2's security level is equivalent to maximizing S_2, and similarly for player 1.

A security dilemma is commonly said to exist when "an increase in one state's security decreases the security of others" (Jervis, 1978, p. 186). It is easy to see that the situation I have just described embodies a security dilemma thus defined, and this is true no matter how "security" is interpreted. For any success by player 2 in decreasing a_1 also decreases b_1, and thereby diminishes player 1's security level. And any success by player 2 in increasing a_2 also increases b_2, and hence increases player 1's subjectively estimated probability that $c_2 < b_2$. Moreover, the same can be said about the relation between player 1's security and player 2's.

A security dilemma thus defined is commonly said to make war more likely, even when no country wants one. Is that true here? It obviously is not. For war will occur in this situation if and only if:

$$b_2 > p_{21}a_2 + (1 - p_{21})c_2 \qquad (3)$$

or an analogous condition exists for player 1. But with the assumptions I have made, that is impossible. Even if one country identifies its security with its security level, continued cooperation will dominate attacking so long as that country does not actually prefer the expected fruits of war to continued cooperation. But if it does, its reason for attacking is that and not the security dilemma.

[6] I assume that with incomplete information about payoffs players assign subjective probabilities to the possible payoffs of other players. In simple finite games with perfect information, this has straightforward implications. In other games the interdependence of players' subjective estimates creates difficulties (as in the next section of this article). For fuller discussion, see Harsanyi (1967-68).

All that the existence of a security dilemma implies in this case is that an increase in one state's confidence that the other does not prefer to attack diminishes the other's confidence that the first prefers not to attack. Thus the subordinate game of selecting optimal retaliatory choices is a zero sum game.

The Incentive to Attack First

Where, then, is the "dilemma" in the security dilemma? In order to find one, we must change some of the assumptions we have made about the relations among the payoffs in Figure 5. Let us continue to assume uncertainty about each other's payoffs on the part of the two players, and also assume that:

$$c_2 > b_2,\ c_1 > a_1$$
$$a_1 > b_1,\ \frac{da_1}{db_1} \geqslant 1 \quad b_2 > a_2,\ \frac{db_2}{da_2} \geqslant 1 \qquad (4)$$
$$\frac{da_1}{da_2} < 0 \quad \frac{db_2}{db_1} < 0$$

In words, let us continue to assume that players 1 and 2 prefer to cooperate rather than attack (although each is uncertain about the other's preference), and that an increase in one security level decreases the other's. But assume now that for each player, the payoff for attacking first is greater than the payoff for retaliating, while positive changes in the latter preserve this inequality.

Now our two players confront two kinds of dilemma. First, it is no longer true that maximizing their security levels will maximize S_1 and S_2, because doing so will tend to increase p_{12} and p_{21}. Thus actions that increase one aspect of their security will diminish another aspect. Second, there is now some danger of military conflict that leaves both players worse off than they would be without it. Because this is a suboptimal equilibrium, this is an outcome analogous to that in the Prisoner's Dilemma. The Prisoner's Dilemma game is not an accurate model of the situation, however. The Stag Hunt comes closer to capturing the problem, but it too is at once an inaccurate and incomplete representation of it.

The truth of these propositions will be demonstrated shortly. Before doing that, let us notice two implications of these propositions. First, the fact that as one state's security increases, another's security decreases is only a necessary and not a sufficient condition for the existence of a genuine security dilemma, no matter which definition of security is meant. But the existence of an incentive to attack first by both countries is a sufficient condition for a security dilemma, and therefore the fact that $\frac{dS_1}{dS_2} < 0$ is a necessary condition is true only in the trivial sense that the former condition implies the latter. The common definition is thus incorrect. Second, the security dilemma (properly defined) has already been subjected to a certain amount of formal analysis. For example, if we apply the payoff orderings in equation (4) to the game matrix in Figure 6, we will see that it is strategically equivalent to the decision-theoretic matrix analyzed by Ellsberg (1960). However, my treatment will differ slightly from Ellsberg's.[7]

Consider now why an incentive to strike first gives rise to the two dilemmas just mentioned. First, we must notice that p_{21} in equation (1) now takes on a different meaning. It is still true that if $a_1 > c_1$, that is a sufficient condition for player 1 to attack. But even if $c_1 > a_1$, player 1 will want to attack if he thinks it sufficiently likely that player 2 will attack (since $a_1 > b_1$). Thus now:

$$p_{21} = q_{21} + (1 - q_{21})r_{21} \qquad (5)$$

where q_{21} is player 2's subjectively estimated probability that $a_1 > c_1$, and r_{21} is player 2's subjectively estimated probability that:

$$a_1 > p_{12}b_1 + (1 - p_{12})c_1 = S_1 \qquad (6)$$

where p_{12} is player 1's subjectively estimated probability that player 2 will attack. It is useful to rearrange equation (6) so that r_{21} becomes player 2's subjectively estimated probability that:

$$p_{12} > \frac{c_1 - a_1}{c_1 - b_1} \qquad (7)$$

Suppose now that player 2 seeks to maximize a_1. Once again the effect is to increase 2's security level and diminish a_1 (and thus q_{21}). Another effect, however, is to *increase* r_{21}, since increasing a_2 will also increase b_2, and hence make player 1 less sure that $c_2 > b_2$. It may also, by decreasing b_1, increase the ratio on the right-hand side of equation (7), and thus increase player 1's sensitivity to doubts he may have about 1's preferences. In the first case we examined, that did not matter (unless one felt sorry for player 1), since player 1 preferred to cooperate whatever the value of S_1. However, with an incentive to attack first, a decrease in S_1 may make $S_1 < a_1$, and hence provoke a preemptive attack by player 1.

It is obvious, then, that an identification of one's "security" with one's security level in this

[7]See also Schelling (1960) and Hunter (1972).

game will maximize the chances of inadvertent conflict. Of course, if the reason for adopting this policy goal is the use of a maximin decision rule, conflict is certain rather than merely likely, since both players maximize their worst outcomes by attacking. However, it is plausible that a government might have the objective of maximizing a_2 or b_1 without following such a decision rule, even though that is inconsistent with evaluating cooperation on the basis of S_2 and S_1. For it may be that this is the organizational mission of its armed services, which the civilian leaders (who act on the basis of S_2 and S_1) are unable to change.

It is also obvious that an identification of one's security with deterrence narrowly construed (i.e., minimizing q_{21}) will maximize the chances of inadvertent conflict in this situation. This, of course, was the point of Ellsberg (1960).

Given a situation such as the one I have just described, what is the rational course of action for players 1 and 2 in selecting retaliatory options? One possible answer to that question is that players 1 and 2 should simply play the zero sum game that results from their each seeking to maximize S_1 and S_2. And if they have no other choices but those among the values of the payoffs they can manipulate, all of which satisfy the conditions given in equation (4), that is what they should do, even though the result might well be suboptimal conflict. For S_1 and S_2 take into account the consequences of their choices for both the likelihood of conflict and the values they will receive should conflict nonetheless occur, and both are relevant.

The problem with this answer is that it is not clear what it implies in any actual situation, since S_1 and S_2 contain implicitly an infinite series of subjective probabilities of subjective probabilities. In order to maximize S_2 with respect to changes in a_2, for example, player 2 must know how q_{21} and r_{21} respond to changes in a_1 and b_2. But this is unlikely to be something that is known with enough precision or confidence to be the basis for a very determinate or generally accepted criterion for policy.[8]

Alternatively, players 1 and 2 can seek to escape from the dilemma by developing options that break one or more of the connections among payoffs listed in equation (4). For example, deterrence by punishment rather than defense breaks the connection between a_1 and a_2 (and b_1 and b_2), whereas reliance on passive defense (for example, fortifications) breaks the connection between a_2 and b_2. In either case it becomes possible for S_2 to be maximized and yet have $b_2 \leqslant a_2$ (and similarly for player 1), and thus the security dilemma (properly defined) no longer exists.

Thus the only alternative to simple (and controversial) maximization of S_1 and S_2 that is consistent with the continued existence of the dilemma in the security dilemma is to try to weaken the connection between each side's first strike payoffs and the other's subjectively estimated probability that one prefers to attack no matter what. The difficulty is that just as single-minded efforts to maximize one's security level are indistinguishable from preparations to attack first, so one's efforts to reassure the other may be interpreted as designed instead to lull him into a false sense of security.

Arms Races

In discussing the security dilemma I have assumed that "cooperate" means "respect the agreement establishing international boundaries," and "defect" means "act to overturn that agreement by force." Decisions about armaments have been analyzed as choices among D alternatives in that game, which affect the payoffs received by both players. We have seen that there is a zero sum component to this subordinate game of defense policy, but that this component alone does not imply a danger of inadvertent war. At this point it is reasonable to ask: How is an arms race ever possible? Given the interdependence between S_1 and S_2, player 2 should expect that any choice he makes to increase S_2 will be countered by player 1. But since player 2 should prefer a given value of S_2 at lower levels of expenditure to the same value at higher levels of expenditure, why would he (or player 1) ever increase defense expenditures? Or, in game theoretic terms, the armaments game is a zero sum game with perfect information. It ought, therefore, to have an equilibrium in pure strategies, and it appears that this equilibrium should be at low levels of expenditure. How, then, are arms races possible?

One possible explanation is that governments are short-sighted and ignore the strategic interdependence involved in defense policy decisions. Each takes the other's armaments level as given and maximizes his own security with respect to it. Such a decision rule in defense policy will produce an arms race, which may or may not have a well-defined equilibrium depending upon the two countries' reaction functions.

There are other possibilities. Governments cannot respond immediately to each other's armaments decisions. Between the time player 1 has in-

[8]Hunter (1972) makes a similar point, although (following Ellsberg) he defines the security objectives of a rational decision maker in such a situation somewhat differently. He also assumes that this point constitutes a telling criticism of deterrence theory, which would be true only if there were some well-defined alternative way of maximizing one's security in such a situation.

creased his arms and player 2 responds, S_1 has been increased at the expense of S_2. If this period is long enough, the increase in S_1 big enough, and player 1's discount rate for future benefits is high enough, player 1 may simply prefer to arm even though he anticipates that this gain will eventually be negated by player 2's response.[9] But player 2 will anticipate this and prefer to arm first himself. The result may be that both are worse off than if neither had armed.

This is a version of the game in Figure 3, in which player 2's response to player 1's defection does not lead to a sufficiently low payoff to 1 to deter him. Although not identical to the Prisoner's Dilemma, this game is similar to it, since even with intersubjective knowledge of utilities there is a suboptimal equilibrium.

It is also possible that, just as the boundary-agreement game contained a security dilemma, so may the armaments game that is subordinate to it. Precisely because governments cannot respond immediately to each other's armaments decisions, each government will prefer to arm first rather than to respond to the other, since its security will be greater in the former case. But even though neither government actually values the temporary increase in its security from an arms increase high enough to prefer that to a continued low level of armaments by both countries, neither may be sure that this is true of the other. Thus we can reinterpret the game in Figure 5 as an armaments game rather than a boundary-agreement game. Variations in payoffs are now the result of advances in technology rather than arms increases. And S_1 and S_2 now refer not to the expected benefits of nonaggression, but to the expected benefits of existing levels of armament. But since the latter must incorporate the former, the problem of interacting subjective estimates is compounded.

Just as the incentive to attack first was a necessary condition for the existence of a security dilemma in the boundary-agreement game, so an incentive to arm first is a necessary condition for either a security dilemma or a straightforward suboptimal equilibrium in the armaments subgame. Other things being equal, then, we would expect longer lead times in weapons production to make arms races more likely.

Let us suppose that $a_1 > S_1$ and $b_2 > S_2$ in the game in Figure 5. Then both sides will want to attack (or, in our alternative interpretation, arm) rather than cooperate. The outcome, however, depends on which one does so first. Thus an incentive to move first makes an arms race or a mobilization race a genuine race (rather than merely an action-reaction process); whoever completes the course first wins. But it is important to recognize that there may be an incentive to attack first, but no incentive to mobilize first or to arm first. If not, and 1 and 2 have not yet mobilized, then the effect of 2's efforts to increase a_2 or diminish a_1 on p_{12} (in the boundary-agreement game) may be quite weak. Thus the extent of the security dilemma cannot be inferred from the mere existence of the conditions stated in equation (4) as applied to the boundary-agreement game.

It should also be clear that Jervis (1978) is incorrect in stating that an advantage to the offense over the defense implies a genuine security dilemma, that is, a danger of war that no one wants. "When we say that the offense has the advantage," he says (1978, p. 187), "we simply mean that it is easier to destroy the other's army and take its territory than it is to defend one's own." Such an advantage implies that improvements in one player's second strike payoffs lead to improvements in his first strike payoffs, and thus that increases in one's security will diminish the other's. But it does not imply that his first-strike payoffs are greater than his second-strike payoffs, for one may have ample time to begin one's own attack after receiving hard evidence that the other has begun his.

It is commonly believed that arms races make wars more likely. Does this analysis provide any support for that belief? The answer depends upon the payoff structure of the boundary-agreement game, and the explanation for the arms race. If the boundary game contains a security dilemma, each government seeks to maximize its security level in it, and each takes the other's arms level as given and reacts to it, then the likelihood of a purely defensive war is great. On the other hand, if the boundary game does not contain a security dilemma, short-sighted maximization of security levels may lead to great insecurity (sometimes for one, sometimes for the other), and wasteful expenditures on armaments, but no war.

What happens if both the boundary game and the armaments game contain security dilemmas? In that case the payoff to player 2 from mutual cooperation in the armaments game (call it c_2*, and similarly for the other payoffs) is simply the value of S_2 at the existing military balance. The payoff to 2 for arming first (b_2*) is the discounted present value of S_2 as player 2 anticipates a temporary arms advantage which is ultimately eliminated by 1's response. But because of the security dilemma in the boundary game, b_2* must take into account the increase in r_{21} (the probability that 1 will launch a preemptive attack) that will result from 2's arms increase. Thus the security dilemma in the boundary game may make both players

[9]Nicholson's (1972) model of oligopoly is directly applicable to this case.

cautious about arms increases and therefore moderate the security dilemma in the armaments game, so long as the initial values of p_{21} and p_{12} are low. Should they become high, however, then c_2^* and c_1^* will be low and the incentive to arm will increase. Thus, with a double security dilemma there may be an unstable equilibrium at low armaments levels, with rapid escalation to an arms race and war should that equilibrium be disturbed.

There is at least one other possible explanation for an arms race that requires neither shortsighted security maximization nor a suboptimal equilibrium of either sort. Precisely because security is costly, a government may respond less fully to another's arms increases at high levels than at low levels. A government that senses a greater willingness than another to pay the costs of defense may rationally believe it can outspend the other and "win" an arms race; player 2, for example, may believe that S_2 will be higher (taking into account 1's predicted response) at high levels of arms expenditures than at low levels. Of course, if that is true, both would be better off if player 2's superiority were acknowledged at lower levels of arms expenditures for both. But with uncertainty, that outcome may not be accepted as inevitable by both, and thus the extra arms expenditures produced by the competition are the costs that both sides paid to find that out.

If there is not an optimal equilibrium in the armaments game, one might reasonably ask whether arms control agreements can be of any help. There are a variety of answers to this question that follow from the above discussion. First, an agreement may serve to make clear to military decision makers the interdependence of their decisions, if that has not been recognized. Second, an optimal equilibrium may exist, but external disturbances may have pushed governments away from it, and no one may have an incentive unilaterally to begin a return to it. Third, governments may cooperate in creating an equilibrium by agreeing to inspection measures, thereby mitigating the problem of reaction time. Finally, an agreement may conceivably help the governments to develop intersubjective knowledge about each other's utilities.

The Terms of Cooperation

Although we dropped some time back the assumption in Prisoner's Dilemma games that players have only two alternatives (C and D), we have so far considered only the implications of their having to choose among a variety of D-type alternatives. It is now time to consider the implications of their ability to alter the values associated with cooperation. Clearly there may be various different forms of cooperation, for which the actors have conflicting preferences. Selection of one agreement over another must then be the result of bargaining. One implication of our discussion of security is that in addition to conflicting evaluations of the terms of various agreements, players may also differ in their evaluations of agreements because of the anticipated levels of security associated with each. Using the notation employed above, with an external enforcer of agreements, players 1 and 2 would evaluate alternative agreements according to c_1 and c_2. Without an enforcer, they will evaluate alternative agreements according to S_1 and S_2. A necessary condition for the acceptance of any agreement is that at the time of the agreement $S_1 > a_1$ and $S_2 > b_2$. However, within the set of agreements satisfying that condition, player 1 will prefer agreements with higher S_1 and player 2 will prefer agreements with higher S_2.

It is common to separate the problem of enforcing agreements from the problem of bargaining over their terms. I now want to suggest that this approach is probably mistaken, not only because (as we have just seen) the necessity for self-enforcement influences actors' evaluations of prospective agreements, but also because actors' relative evaluations of the difference between agreement and nonagreement will affect which agreements are enforceable.

In discussing this question we are hampered by the lack of consensus on the best way to understand the bargaining process. However, a feature common to a number of different bargaining theories is that bargainers are influenced not simply by their evaluations of possible agreements, but also by their evaluations of the consequences of nonagreement. (This, of course, is the basis for the influence of threats.)

But this implies that for an agreement to be stable, it must be possible not only for its adherents to be able credibly to threaten to match each other's noncooperation with their own, but also that the two players' evaluations of noncooperation continue to be such as to sustain the *terms* of the agreement that emerged from the bargaining process that led to it. Otherwise the balance of bargaining strength will have shifted, and one or the other will have an incentive to threaten to withdraw his cooperation until the terms of the agreement are modified. Thus the possibility that the consequences of responding to a defection after an agreement has been implemented may be worse than a failure to accept the agreement in the first place means that such agreements may not be acceptable, even when there is intersubjective knowledge of utilities, for their result may be to alter the relative bargaining power of the adherents.

Shubik (1970) has already stressed the impor-

tance of a theory of threats for understanding decentralized enforcement of cooperation. My point here is that a theory of threats requires a theory of bargaining as well. Any threat that can be used to enforce one agreement as opposed to another can also be used to enforce adherence to that agreement, as long as the participants maintain the ability to carry out their threats throughout the life of the agreement. But a threat that *cannot* enforce one agreement as opposed to another cannot be used to enforce adherence to it.

Earlier I assumed that a player will overturn an agreement, including agreements establishing international boundaries, only if he prefers the expected value of nonagreement (for example, the fruits of military conquest) to the continuation of the agreement, or, in the case of the security dilemma, if he fears that the other is about to defect. We have just seen that once the possibility of bargaining is introduced, there is another possible motivation: a player may defect as a means of forcing the other to agree to an alteration in the terms of the agreement.

But the result in all three cases may be negotiation rather than war. Suppose that player 1 in Figure 5 comes to prefer a_1 to c_1. Player 2 may nonetheless avoid the costs of conflict by conceding something to player 1 (for example, a portion of the territory he might have captured in war), such that both 1 and 2 prefer this new agreement to the consequences of conflict. Since both players 1 and 2 should normally prefer the outcome of military conflict without the costs of war to the same outcome with the costs of war, why would they fight rather than agree to this alteration in c_1 and c_2? One possible reason is that they disagree about the outcome to be expected as the result of conflict.

But exactly the same can be said of the outbreak of conflict as a bargaining strategy. If the outcome of bargaining can be anticipated in advance, it is in the interests of both parties to agree to it immediately. If they do not, a plausible explanation is that they do not agree about what outcome will result from the actual execution of threats.

Incomplete information about utilities, then, implies a significant probability of conflict for reasons that have nothing to do with the security dilemma or the lack of enforceability of agreements. Instead of asking, "Why do wars occur?" we might ask, "Why do strikes occur?" The questions have a similar motivation: frequently in both cases no one gains enough to compensate for the cost of conflict. In the case of strikes, the explanation cannot be the unenforceability of agreements. A plausible one (consistent with much bargaining theory) is that if the outcome could have been confidently predicted in advance, it would have been agreed to immediately, but because the parties had divergent expectations of the outcome, they each chose to put them to the test. The effect of government on such conflicts is to limit the kinds of damage that the parties to them might try to inflict on each other.

Because actors may make concessions as a means of avoiding the costs of conflict, there is an incentive to bluff—to exaggerate one's own expectations about the outcome of war, or to exaggerate one's own bargaining power. In situations characterized by the security dilemma that is obviously dangerous, since the result may be to provoke not a concession, but a preemptive attack.

In the preceding discussion of security, security dilemmas, and bargaining, I have mainly interpreted "cooperation" to mean acceptance of a set of international boundaries, and "defection" as the use of force to overturn them. However, the implications of this analysis are obviously much more general than that. For example, everything I have said applies to Hirschman's (1945) analysis of the risks associated with foreign trade.

The problem of arms control includes a special complication that should be explicitly recognized. Most international agreements can easily be interpreted as outcomes that leave both participants better off, but perhaps one fares better than the other because of some bargaining advantage. But an arms control agreement is an agreement in which the participants jointly regulate their abilities to threaten each other, and therefore to derive a bargaining advantage in other contexts.

At first glance, it seems reasonable to suggest that they will therefore only agree to arms control arrangements that leave their relative abilities to bargain unchanged but reduce the costs to each of sustaining those abilities. If true, this in itself helps us understand one problem in reaching such agreements. It will normally not be intersubjectively obvious when two countries' relative bargaining strengths have or have not been influenced by a proposed arms reduction. Moreover, if the point of normal bargaining is to *demonstrate* one's superior strength by actually applying the threats one has made, this method of settling disagreements is not available in arms control negotiations. (One does not go to war to demonstrate that one could *not* have prevailed under the terms of the agreement.)

But there may be more to the problem; the point of the exercise is that maintaining a capacity for making military threats entails a burden even when they are not used. But this implies that differences in willingness to bear that burden can lead indirectly to a bargaining advantage in the military bargaining game itself. Thus governments may not be satisfied with arms control agreements that leave relative military bargaining power in-

tact. They may also seek a ratification of any superiority they sense in their willingness to bear the burdens of armaments in peacetime. But as we have already seen, this can in itself lead to an arms race, as governments try to enforce on each other lesser degrees of security by competitive demonstrations of their relative willingness to bear this burden.

n-Person Games

Let us now consider the remaining unaltered condition assumed by the standard Prisoner's Dilemma, that there are only two players. As long as the other conditions also hold, this is not a significant limitation, although a new complication does have to be considered: the payoffs to each player may now be a function of the *number* of other players who cooperate (Schelling, 1973). However, if the conditions already discussed are relaxed, then significant additional problems emerge.

First, of course, if players have more than two alternatives they may be able to consider not only the terms on which they will cooperate, but also with whom they will cooperate. Thus all the complexities of *n*-person game theory must be introduced, although without the assumption of the theory of cooperative games that any mutually desirable coalition can form.

However, even if we ignore that set of problems and assume that the identity of those who would benefit from cooperating is fixed, the complications already discussed take on added complexities in the *n*-person case. First, of course, the problem of settling the terms of cooperation becomes much more complex. As a result, as the number of cooperators increases, the time required to negotiate the terms of their cooperation increases as well. If we assume that individuals discount future benefits, then a new barrier to cooperation emerges: the present value of any future agreement may be less than nonagreement, and cooperation no longer is Pareto optimal. Thus the problem of reducing decision costs while settling conflicts of interest among potential cooperators emerges as an important problem in organizing cooperation among more than two actors.

Second, if the underlying preferences really are those of the Prisoner's Dilemma, then cooperation can be an equilibrium only if it is conditional. (This is true whether the situation is a version of the one presented in Figure 3, or a Prisoner's Dilemma supergame.) But this requires that each decides in full knowledge of decisions made by the others, and that no one be able to count on having the last word. As the number of potential cooperators increases, the satisfaction of both these conditions may become more difficult (Taylor, 1976, pp. 92-93).

For two reasons, however, this does not imply that with large numbers, the Prisoner's Dilemma reemerges as an accurate portrayal of the problem of international cooperation. First, the actual emergence of a large-scale social organization is surely not unrelated to the ability of small groups to dominate large groups. That it may be harder to organize large groups than small groups may be an important part of the explanation for that ability. Second, insofar as the major states that exist behave as unitary actors, cooperation among them can often be arranged, since relations among them often satisfy the conditions necessary for conditional cooperation.

Conclusions

A central analytical problem in the study of international relations is to identify under what circumstances the decentralized enforcement of agreements is rational. An answer to that problem should help explain why there is so much order in the anarchic world of international relations, but no more. Often the preferences of the players in the well-known Prisoner's Dilemma game seem to represent the preferences of persons or governments confronted with this problem. I have shown that nonetheless, neither the simple Prisoner's Dilemma game nor the supergame based on it is as helpful in thinking about it as is commonly supposed.

I have pointed out a number of ways in which the Prisoner's Dilemma game fails to model accurately many of the situations in international relations to which it has been applied, even when the preferences of the participants are identical to the preferences assumed in that game. Among them is the assumption implicit in that game that the players act in ignorance of each other's choices, and that each chooses only once. In many situations the participants can respond to each other's choices, and no one can count on having the last choice. Whenever that is true, conditional cooperation can be an equilibrium outcome, even though the players' preferences are those of the Prisoner's Dilemma, and the game is played only once. This contradicts the prevalent notion that, when individuals' preferences are those of the Prisoner's Dilemma, conditional cooperation is rational only if the game is indefinitely repeated (Taylor, 1976, pp. 104-105).

Even when this point has been recognized (as by Snyder), its implications have been missed by continuing to work with the Prisoner's Dilemma matrix. Explicit consideration of a wider range of strategies (even when the players have the same number of choices), allows one, for example, to

develop a game theoretic model of the security dilemma. Such a model demonstrates that the security dilemma, as commonly defined, need not have the implications commonly ascribed to it. Fortunately, the fact that an increase in one nation's security must diminish another's is not a sufficient condition for the existence of a danger of a purely defensive conflict, or an arms race.

Thus theories of the security dilemma, like those based on the Prisoner's Dilemma, lead to excessively pessimistic inferences from the preferences attributed to governments. The same is true of Jervis's game theoretic representation of Rousseau's Stag Hunt, which assumes that the problem of international cooperation is a game of imperfect information, and therefore greatly exaggerates the problem of achieving jointly preferred outcomes when there is incomplete knowledge of preferences.

A more general conclusion to be drawn from all this is that any game matrix can accurately represent a situation only if it includes all the possible strategies. But strategies depend not simply on the alternatives available to the players, but also on the sequence in which they choose and what information they have when choosing. Thus only very simple situations can be adequately modeled without representation first in extensive form. (And even then the normal form can be misleading.) (Harsanyi, 1978, pp. 53-60) More careful attention to this basic point would show that none of the classic 2 × 2 games can do the work that is so often assigned to them.[10]

Finally, the analysis offered bears on the question of the effects of anarchy on the relations among states. Use of the Prisoner's Dilemma (or the Stag Hunt) as either a metaphor or a model in discussions of that question has the effect of focusing our attention on the suboptimality of outcomes that can result when agreements cannot be enforced. Although my analysis indicates that this is certainly a problem, it also indicates that its extent may have been exaggerated.[11]

In addition to this problem, however, there is another effect of anarchy that should be emphasized: collective decisions at the global level are made by bargaining rather than by voting. Thus some of the attributes of international relations that have been attributed to the absence of an enforcer of agreements are really the effect of the costliness (and possible unfairness) of bargaining as a means of making collective decisions (Young, 1978).

The difficulties arising from both these attributes of anarchic systems are obviously compounded by a third: the absence of externally imposed constraints on the actions that may be taken by the most powerful actors.

[10]Snyder and Diesing (1977, p. 58) say of the extensive form, "We did not find this model to be useful." As a result, any resemblance between the game matrices they discuss and the crises examined in their book is purely accidental.

[11]See also Young (1979).

References

Axelrod, R. The emergence of cooperation among egoists. *American Political Science Review,* 1981, *75,* 306-318.

Brams, S. J., David, M. D., & Straffin, P. D. The geometry of the arms race. *International Studies Quarterly,* 1979, *23,* 567-588.

Ellsberg, D. *The crude analysis of strategic choices.* RAND P-2183. Santa Monica, Calif.: Rand Corporation, 1960.

Harsanyi, J. C. Games with incomplete information played by "Bayesian" players, parts 1-3. *Management Science,* 1967-1968, *14,* 159-182; 320-334; 486-502.

Harsanyi, J. C. A solution theory for noncooperative games and its implications for cooperative games. In P. C. Ordeshook (Ed.), *Game theory and political science.* New York: New York University Press, 1978.

Hirschman, A. *National power and the structure of foreign trade.* Berkeley: University of California Press, 1945.

Howard, N. *Paradoxes of rationality: theory of metagames and political behavior.* Cambridge, Mass.: MIT Press, 1971.

Hunter, D. E. Some aspects of a decision-making model in nuclear deterrence theory. *Journal of Peace Research,* 1972, *3,* 209-222.

Jervis, R. Cooperation under the security dilemma. *World Politics,* 1978, *30,* 167-214.

Nicholson, M. *Oligopoly and conflict.* Liverpool: Liverpool University Press, 1972.

Schelling, T. The reciprocal fear of surprise attack. In *The strategy of conflict.* Cambridge, Mass.: Harvard University Press, 1960.

Schelling, T. Hockey helmets, concealed weapons, and daylight saving: a study of binary choices with externalities. *Journal of Conflict Resolution,* 1973, *17,* 381-428.

Sen, A. Rational fools: a critique of the behavioral foundations of economic theory. *Philosophy and Public Affairs,* 1977, *4,* 317-344.

Shubik, M. Game theory, behavior, and the paradox of the prisoner's dilemma: three solutions. *Journal of Conflict Resolution,* 1970, *14,* 181-193.

Snyder, G. H. "Prisoner's dilemma" and "chicken" models in international politics. *International Studies Quarterly,* 1971, *15,* 66-103.

Snyder, G. H., & Diesing, P. *Conflict among nations: bargaining, decision making, and system structure in international crises.* Princeton, N.J.: Princeton University Press, 1977.

Taylor, M. *Anarchy and cooperation.* New York: Wiley, 1976.

von Neumann, J., & Morgenstern, O. *Theory of games*

and economic behavior. Princeton, N.J.: Princeton University Press, 1944.

Wagner, R. H. Deterrence and bargaining. *Journal of Conflict Resolution,* 1982, *26,* 329-358.

Young, O. Anarchy and social choice: reflections on the international polity. *World Politics,* 1978, *30,* 241-263.

Young, O. *Compliance and public authority: a theory with international applications.* Baltimore: Johns Hopkins University Press, 1979.

Young, O. International regimes: problems of concept formation. *World Politics,* 1980, *32,* 331-356.