Newcomb's Problem and Prisoners' Dilemma
Author(s): Steven J. Brams
Source: *The Journal of Conflict Resolution*, Vol. 19, No. 4 (Dec., 1975), pp. 596-612
Published by: Sage Publications, Inc.
Stable URL: http://www.jstor.org/stable/173326
Accessed: 18/11/2013 08:26

# Newcomb's Problem and Prisoners' Dilemma

STEVEN J. BRAMS
*Department of Politics*
*New York University*

The relationship between Newcomb's problem, which involves an apparent paradox of prediction, and Prisoners' Dilemma is explicated. After describing a resolution to Newcomb's problem, due to John A. Ferejohn, that renders the two contradictory principles of choice in Newcomb's problem (dominance and expected utility) consistent, I show Prisoners' Dilemma to be a "symmetricized" version of Newcomb's problem in its payoff structure. The assumption about predictability of choice made for one player in Newcomb's problem, when applied to both players in Prisoners' Dilemma—one considered as a leader and the other as a follower—offers a resolution to this dilemma that, while consistent with the resolution offered by metagame theory, rationalizes the existence of a metagame solution within a probabilistic, rational-choice framework. The relevance of the mutual-predictability assumption to the solution of arms races, and tragedy-of-commons situations generally, is discussed.

Game theory is rife with paradoxes, the most famous of which is illustrated by the game of Prisoners' Dilemma (Rapoport and Chammah, 1965; Rapoport, 1974). Several years ago Anatol Rapoport (1967) suggested that one could escape this paradox through the use of metagame theory, which is a theory developed by Nigel Howard (1971, 1974) that allows players to make successive predictions about each other's condi-

AUTHOR'S NOTE: This paper was presented at the Thirteenth North American Conference, Peace Science Society (International), Cambridge, Massachusetts, November 10-12, 1975.

JOURNAL OF CONFLICT RESOLUTION, Vol. 19 No. 4, December 1975

tional strategy choices.[1] More recently, an apparently unrelated paradox of prediction, posed by William A. Newcomb and elucidated by Robert Nozick (1969; see also Bar-Hillel and Margalit, 1972; Schlesinger, 1974; Levi, 1975), generated a huge response from *Scientific American* readers after being discussed by Martin Gardner (1973). Nozick, who responded to the correspondents several months later, concluded that "the letters do not, in my opinion, lay the problem to rest" (Gardner, 1974: 108).

In this essay I shall show that there is an intimate connection between Newcomb's problem and Prisoners' Dilemma, the latter being a "symmetricized" version of the former in its payoff structure. Moreover, I shall show that an assumption made about one player in Newcomb's problem, when applied to both players in Prisoners' Dilemma—one considered as a leader and the other as a follower—offers a resolution to this dilemma that is generally consistent with the resolution offered by metagame theory. Unlike metagame theory, however, the solution proposed is not based on the *assumption* that players can successively predict each other's strategy choices, before the game is played, but rather is derived as a *consequence* of calculations that maximize the players' expected utility. I shall also describe a persuasive (and separate) resolution to Newcomb's problem due to John A. Ferejohn.

## NEWCOMB'S PROBLEM

Imagine the following situation. There are two boxes, B1 and B2. B1 contains $1,000; B2 contains either $1,000,000 or nothing, but you do not know which. You have a choice between two possible actions:

(1) Take what is in both boxes.

(2) Take only what is in B2.

Now what is in B2 depends on what action some superior Being predicted you would take beforehand. If he predicted you would (1) take what is in both boxes—or would randomize your choice between the two actions—he put $0 in B2; if he predicted you would (2) take only what is in B2, he put $1,000,000 in B2. Hence, you are rewarded for taking only what is in

[1] There has been considerable controversy over the metagame solution to Prisoners' Dilemma and other two-person, non-zero-sum games. For a discussion of, and references to, the different viewpoints that have been expressed, see Brams (1975b: 30-50). Other "solutions" to Prisoners' Dilemma have been proposed by Shubik (1970) and Hill (1975), but none seems as relevant to the study of political conflict as does the theory-of-metagames solution.

B2—provided the Being predicted this choice—though you have some chance of getting even more ($1,001,000) if you take what is in both boxes and the Being incorrectly predicted that you would take only what is in B2. On the other hand, you do much less well ($1,000) if you take what is in both boxes—and the Being predicted this action—and worst ($0) if you take what is in B2 and the Being incorrectly predicted that you would take what is in both boxes.

These payoffs are summarized in the payoff matrix of Figure 1. Clearly, the very best ($1,001,000) and very worst ($0) outcomes occur when the Being's predictions are incorrect, the intermediate outcomes ($1,000,000 and $1,000) when the Being's predictions are correct.

Note that the Being's strategies given in Figure 1 are predictions, not what he puts in B2. We could as well define his two strategies to be "Put $1,000,000 in B2" and "Put $0 in B2," but since these actions are in one-to-one correspondence with his predictions about what you take, it does not matter whether we consider the Being's strategies to be predictions or actions. (Since the Being's predictions precede his actions, they are perhaps the more basic indicator of his behavior.)

From the perspective of the game theorist, what does matter is that the Being's strategies are not the "free" choices usually assumed of players in the normal-form representation of a game. But this is not a game in the usual sense, which renders its Figure 1 representation vulnerable to attack. Moreover, the solution I shall propose to a symmetrical version of this game rests on a different model of player choices.

On first blush, it would appear, Newcomb's problem does not present you with a problem of choice. Your second-row strategy—to take what is in both boxes—dominates your first-row strategy—to take only what is in B2—since whatever the Being predicts, your payoffs are greater than those in your first row. Thus, you should always take what is in both boxes,

| | | *Being* | |
|---|---|---|---|
| | | Predicts you take only what is in B2 | Predicts you take what is in both boxes |
| *You* | Take only what is in B2 | $1,000,000 | $0 |
| | Take what is in both boxes | $1,001,000 | $1,000 |

**Figure 1: Payoff Matrix for Newcomb's Problem**

which assures you of at least $1,000, as contrasted with a minimum of $0 for your first-row strategy.

This choice is complicated, however, by your knowledge of the past performance of the Being, who is (or seems) "superior" precisely because his predictions have always been correct in the past. Although you do not know what his prediction is in the above choice situation, it will be, you believe, almost surely correct. Thus, if you choose your dominant strategy of taking what is in both boxes, the Being will almost surely have anticipated this action and have left B2 empty. Hence, you will get only $1,000 from B1.

On the other hand, if you choose your first-row strategy and take only what is in B2, the Being, expecting this, will almost surely have put $1,000,000 in B2, which would seem a strong argument for your choosing this strategy, despite the dominance of your second-row strategy. This argument is based on the principle of maximizing "expected utility," which is the sum of the payoffs associated with each of the mutually exclusive outcomes in each row times the probability that each will occur. In the example, if the probability that the Being is correct is greater than 0.5005, the expected utility of your first-row strategy will exceed that of your second-row strategy.

This conflict between the dominance principle, which prescribes taking what is in both boxes, and the expected-utility principle, which prescribes taking only what is in the second box, is the heart of the paradox. Is there any solution to this paradox that resolves the conflict between these two principles?

## WHICH PRINCIPLE, AND IS THERE A CONFLICT?

John A. Ferejohn has shown that if Newcomb's problem is reformulated as a *decision-theoretic* rather than as a *game-theoretic* problem, the apparent inconsistency between the two principles disappears.[2] Conceptualized in these terms, the person making the choice of either B2 or both boxes in Newcomb's problem does not view the Being as making

[2] Personal communication, John A. Ferejohn, May 27, 1975. Nigel Howard has also shown these two principles to be consistent in a metagame representation of Newcomb's problem (Personal communications, March 27, 1975 and June 25, 1975). Whereas Howard's metagame resolution of the paradox retains the assumption that Newcomb's problem is a game, Ferejohn criticizes precisely this assumption, as I indicate in the text. I am grateful to both scholars for their comments on earlier versions of this paper, but neither should be held responsible for the conclusions that I draw from the present analysis.

predictions *about what he will choose,* but rather as making predictions *that are correct or incorrect* (see Figure 2).

Recall that your two best outcomes in the payoff matrix of Figure 1 ($1,000,000 and $1,001,000) were both associated with the Being's predicting that you would take only what is in B2 (first column of Figure 1). In the decision-theoretic payoff matrix of Figure 2, by contrast, these outcomes are the diagonal elements, each being associated with a different "state of nature," which is assumed to be either a correct or an incorrect prediction on the part of the Being. Because your best choice depends on what state of nature obtains in the decision-theoretic representation (if the Being is correct, take only what is in B2; if the Being is incorrect, take what is in both boxes), neither of your two actions dominates the other.

Since you do not have a dominant strategy in the decision-theoretic representation of Figure 2, there no longer exists a conflict between the expected-utility principle and the dominance principle. Now the sole determinant of whether you should take only what is in B2, or you should take what is in both boxes, to maximize your expected utility are the probabilities that you associate with each state of nature. If the probability that the Being is correct is greater than 0.5005, then you should take only what is in B2; if this probability is less than 0.5005, then you should take what is in both boxes; and if this probability is exactly 0.5005, then you would be indifferent between your two actions.

How persuasive is this resolution of Newcomb's problem? If you believe that the Being has no control over which state of nature obtains in Figure 2, then the Being is not properly a player in a two-person game of the kind assumed in Figure 1; hence, the appropriate representation of Newcomb's problem is decision-theoretic. To be sure, the probabilities of being in each state are not specified by Newcomb's problem, so the decision-theoretic representation does not answer the question of whether you should take

| | | *State of nature* | |
|---|---|---|---|
| | | Being correct | Being incorrect |
| *You* | Take only what is in B2 | $1,000,000 | $0 |
| | Take what is in both boxes | $1,000 | $1,001,000 |

**Figure 2: Newcomb's Problem as a Decision-Theoretic Problem**

only what is in B2 or whether you should take what is in both boxes. However, this representation does demonstrate that there is no conflict between the dominance principle and the expected-utility principle.

On the other hand, if you believe that the Being has some control over which state of nature obtains—which is a question quite different from whether he can predict your choice (which he almost surely can)—then he is not an entirely passive state of nature, at least with respect to being correct; hence, the game-theoretic representation of Figure 1 is the appropriate one. However, it must be said that there is nothing in the original statement of Newcomb's problem to indicate that the Being's choices are anything but mechanistic—that is, the correctness of his prediction about your action is *not* assumed to depend in any way on your choice. Or, to put it another way, though you are assumed to exercise free will with respect to the action you take, the Being exercises no free will with respect to what he puts in B2; his "choice" is dictated solely by his prediction.

The fact that the Being's prediction is assumed to be almost surely correct would seem to imply that you are indeed playing a game against nature whose two states—Being correct or Being incorrect—occur with the same relative frequency whatever you do. Given that this is the proper interpretation of Newcomb's problem, then Ferejohn's ingenious decision-theoretic reformulation of the problem convincingly resolves the presumed conflict between the dominance and expected-utility principles.

## NEWCOMB'S PROBLEM SYMMETRICIZED: PRISONERS' DILEMMA

If we can dispose of Newcomb's problem in the above manner, it is still intriguing to ask what consequences the predictive ability assumed on the part of the Being would have if *both* actors in Newcomb's problem could make genuine choices as players in a game. We may generalize the payoff matrix of Newcomb's problem to that shown in Figure 3, where the payoffs in the matrix represent *utilities* of the outcomes to the row player (A), $A_1$ representing his best payoff, $A_2$ next, and so on. The dominance principle says that player A should choose strategy $a_2$, the expected-utility principle says that player A should choose strategy $a_1$, given that A considers B's ability to predict his (A's) choices to be "sufficiently good." More precisely, if p is the subjective probability that A believes B's prediction about his strategy choice will be correct, then the expected-utility principle would prescribe that A should choose strategy $a_1$ if

$$A_2p + A_4(1 - p) > A_1(1 - p) + A_3p.$$

In Newcomb's problem, an asymmetry is assumed in both the abilities and actions of the two players in the prediction-choice game. The Being (player B in Figure 3) is assumed to be a phenomenally good guesser, but no such superior intelligence is attributed to the chooser (player A in Figure 3). Furthermore, player B is assumed to make the first move, but in fact this gives him neither an advantage nor a disadvantage because his choice of what to put in the boxes (based on his prediction) is not communicated to player A. Thus, we could just as well assume that the two players make simultaneous choices; the essential nature of the game remains unchanged.

The game does change, however, if we assume not only that player B can make predictions about player A's choices, but also that A can make predictions about B's choices as well. If player B's ranking of the outcomes duplicates player A's in Figure 3–but now, with the rows and columns interchanged, A is assumed to be the predictor and B the chooser–the payoff matrix of player B will appear as in Figure 4, with $B_1$ representing his best payoff, $B_2$ next, and so on. As with player A in the Figure 3 game, the dominance principle and the expected-utility principle prescribe different strategy choices for player B if he (B) considers player A's ability to predict his choices to be sufficiently good.

If we combine the payoffs in the two asymmetrical prediction-choice games into a single payoff matrix, we get the game shown in Figure 5 (in which only the players' strategies, but not their predictions about the other player's strategy choices, are shown). The payoff matrix for this game gives the outcomes for both players, where, for each cell entry $(A_i,B_j)$, $A_i$ represents the payoff to the row player and $B_j$ the payoff to the column player. The ranking of outcomes by both players in this game

| | | *Player B* | |
|---|---|---|---|
| | | Predicts $a_1$ | Predicts $a_2$ |
| *Player A* | $a_1$ | $A_2$ | $A_4$ |
| | $a_2$ | $A_1$ | $A_3$ |

**Figure 3: Generalized Payoff Matrix for Player A in Newcomb's Problem**

| | | *Player B* | |
|---|---|---|---|
| | | $b_1$ | $b_2$ |
| | Predicts $b_1$ | $B_2$ | $B_1$ |
| *Player A* | | | |
| | Predicts $b_2$ | $B_4$ | $B_3$ |

**Figure 4: Payoff Matrix for Player B**

generated by the symmetric play of Newcomb's prediction-choice game defines the classic 2 x 2 Prisoners' Dilemma game.

The dilemma for the players in this game lies in the fact that whereas they both prefer outcome ($A_2$, $B_2$) to outcome ($A_3$, $B_3$), the former outcome is not in equilibrium: each player has an incentive to shift to his second (dominant) strategy, given that the other player sticks to his first (dominated) strategy. Both players are therefore motivated to "play it safe" and to choose their dominant second strategies ($a_2$ and $b_2$), which—unfortunately for them—yields the "noncooperative" outcome ($A_3$, $B_3$) that both find inferior to the "cooperative" outcome ($A_2$, $B_2$).

## A SOLUTION TO PRISONERS' DILEMMA

The fact that the problems of choice in Newcomb's problem and Prisoners' Dilemma are related should not obscure the fact that the latter is a two-person game—in which both players can make free and independent choices—whereas the former seems best conceptualized as a (one-person) game against nature, or a situation of decision-making under risk. However, the condition in the symmetric version of Newcomb's

| | | *Player B* | |
|---|---|---|---|
| | | $b_1$ | $b_2$ |
| | $a_1$ | ($A_2$,$B_2$) | ($A_4$,$B_1$) |
| *Player A* | | | |
| | $a_2$ | ($A_1$,$B_4$) | ($A_3$,$B_3$) |

**Figure 5: Combined Payoff Matrix for Players A and B**

problem that each player knows that the other player can predict—with a high degree of accuracy—which strategy he will choose does have a surprising consequence for the play of Prisoners' Dilemma: it provides an incentive for each player *not* to choose his second dominant strategy ($a_2$ or $b_2$).

True, if one player knows that the other player will almost surely choose his second strategy, then he will also choose his second strategy in order to insure against receiving his worst payoff ($A_4$ or $B_4$). As a consequence of these choices, the noncooperative outcome ($A_3$, $B_3$) will be chosen.

But now assume that one player knows that the other player plans—at least initially—to select his first strategy. Then one would ordinarily say that he should exploit this information and select his second strategy, thereby realizing his best payoff ($A_1$ or $B_1$). But this tactic will not work, given the mutual predictability of choices we have assumed on the part of both players in this symmetric version of Newcomb's problem. For any thoughts by one player of "defecting" from his strategy associated with the cooperative but unstable outcome, ($A_2$, $B_2$), would almost surely be detected by the other player. The other player then could exact retribution—and at the same time prevent his worst outcome from being chosen—by switching to his own noncooperative strategy. Thus, the mutual predictability of strategy choices that we have assumed in the symmetric version of Newcomb's problem helps to ensure against noncooperative choices by either player and to stabilize the cooperative solution to Prisoners' Dilemma.

## THE EXPECTED-UTILITY ARGUMENT

More formally, assume player A contemplates choosing either strategy $a_1$ or $a_2$ and knows that player B can correctly predict his choice with probability p and incorrectly predict his choice with probability $1 - p$. Similarly, assume that player B, facing the choice between strategy $b_1$ and $b_2$, knows that player A can correctly predict his choice with probability q and incorrectly predict his choice with probability $1 - q$. Given these probabilities, I shall now show that there exists a "choice rule" that *either* player can adopt that will induce the other player to choose his cooperative strategy—based on the expected-utility criterion—given that the probabilities of correct prediction are "sufficiently high."[3]

[3] For a different concept of inducement, based on players' misrepresentation of their preferences in 2 x 2 games, see Brams (1975a).

A *choice rule* is a conditional strategy based on one's prediction about the strategy choice of the other player. In the calculation to be given below, we assume that one player adopts a choice rule of *conditional cooperation:* he will cooperate (i.e., choose his first strategy) if he predicts that the other player will also cooperate by choosing his first strategy; otherwise, he will choose his second (noncooperative) strategy.

Assume that player B adopts a choice rule of conditional cooperation. Then if player A chooses strategy $a_1$, B will correctly predict this choice with probability p and hence will choose strategy $b_1$ with probability p and strategy $b_2$ with probability $1 - p$. Thus, given conditional cooperation on the part of B, A's expected utility from choosing strategy $a_1$ will be

$$E(a_1) = A_2p + A_4(1-p).$$

Similarly, his expected utility from choosing strategy $a_2$ will be

$$E(a_2) = A_1(1-p) + A_3p.$$

Comparing $E(a_1)$ and $E(a_2)$,

$$A_2p + A_4(1-p) \overset{?}{>} A_1(1-p) + A_3p,$$

$$(A_2 - A_3)p \overset{?}{>} (A_1 - A_4)(1-p),$$

$$p/(1-p) \overset{?}{>} (A_1 - A_4)/(A_2 - A_3),$$

we see that this inequality is satisfied, and $E(a_1) > E(a_2)$, when p (in comparison to $1 - p$) is "sufficiently large." If, for example, the utilities associated with player A's payoffs are $A_1 = 4$, $A_2 = 3$, $A_3 = 2$, and $A_4 = 1$, then the expected utility of player A's first strategy will be greater than that of his second strategy if

$$p/(1-p) > (4-1)/(3-2),$$

$$p > 3(1-p),$$

$$4p > 3,$$

$$p > 3/4.$$

That is, by the expected-utility criterion player A should choose his first (cooperative) strategy if he believes that player B can correctly predict his strategy choice with a probability greater than ¾, given that player B responds in a conditionally cooperative manner to his predictions about A's choices. Note that whatever the utilities consistent with player A's ranking of the four outcomes are, p *must* exceed ½.

What happens if player B adopts a less benevolent choice rule? Assume, for example, that he always chooses strategy $b_2$, whatever he predicts about the strategy choice of player A. Then if A now adopts a conditionally cooperative choice rule, he will choose strategy $a_1$ with probability $1 - q$ and strategy $a_2$ with probability q. Thus, given conditional cooperation on the part of A, B's expected utility from always choosing strategy $b_2$ will be

$$E(b_2) = B_1 (1 - q) + B_3 q.$$

Similarly, his expected utility from always choosing strategy $b_1$ will be

$$E(b_1) = B_2 q + B_4 (1 - q).$$

Comparing $E(b_1)$ and $E(b_2)$, we can show, in a manner analogous to the comparison of strategies given for player A earlier, that $E(b_1) > E(b_2)$ if

$$q / (1 - q) > (B_1 - B_4) / (B_2 - B_3),$$

i.e., whenever q (in comparison to $1 - q$) is "sufficiently large." Subject to this condition, therefore, player B would *not* be well advised always to choose strategy $b_2$ if player A adopts a conditionally cooperative choice rule. Clearly, a choice rule of noncooperation on the part of one player is inconsistent with a choice rule of conditional cooperation on the part of the other player.

## COOPERATION OR NONCOOPERATION?

So far I have shown that if one player—call him the *leader*—(1) adopts a conditionally cooperative choice rule and (2) can predict the other player's strategy choice with a sufficiently high probability, the other player—call him the *follower*—maximizes his own expected utility by also cooperating, given that he can detect lies on the part of the leader with a sufficiently high probability. Thereby both players "lock into" the cooperative solution, which—it will be recalled—is unstable in Prisoners' Dilemma when the players do not have the ability to predict each other's strategy choices.

There is one question that remains, however. Given that the follower maximizes his expected utility by cooperating when the leader adopts a choice rule of conditional cooperation, how does the follower know when the leader adopts such a choice rule in the first place? The answer is that he does not, unless the leader announces his intention to adopt this choice rule.

To escape the dilemma, therefore, we must assume that there is some communication between the players. Moreover, we must assume that one player (the leader) announces a choice rule to which the other player (the follower) responds. If neither player takes the initiative, nothing can happen; if both players take the initiative simultaneously and announce the choice rule of conditional cooperation, each presumably will await a commitment on the part of the other before committing himself, and nothing again will happen. Should the players simultaneously announce different choice rules, the resulting inconsistencies may lead to confusion, or possibly an attempt to align the rules or distinguish the roles of leader and follower.[4]

The only clean escape from the dilemma, therefore, occurs when the two players can communicate and take on the distinct roles of leader and follower. Although, strictly speaking, permitting communication turns Prisoners' Dilemma into a game that is no longer noncooperative, communication alone is not sufficient to resolve the dilemma without mutual predictability. For what is to prevent the leader from lying about his announced intention to cooperate conditionally? And what is to prevent the follower from lying about his announced response to select his cooperative strategy?

The insurance against lies that players have with mutual predictability is that the lies can be detected with probabilities p and q. If these probabilities satisfy the previous inequalities, then it pays for the follower to cooperate in the face of a choice rule of conditional cooperation, and for the leader to cooperate by then choosing his cooperative strategy, too. Otherwise, the insurance both players have against lying will not be sufficient to make cooperation worth their while, and they should choose, instead, their noncooperative dominant strategies. We conclude, therefore, that a mutual ability to predict strategy choices on the part of both players offers them a mutual incentive to choose their cooperative strategies.

[4] The so-called Stackelberg solution in duopoly theory in economics also distinguishes between a leader and a follower (Henderson and Quandt, 1971: 229-231).

## RELATIONSHIP TO METAGAME SOLUTION

The solution to Prisoners' Dilemma proposed here has similarities to the solution of this game prescribed by metagame theory, but there are also some significant differences. In this theory, the successive iteration of conditional strategies by the players yields some strategies in the end at whose intersection the cooperative outcome is in equilibrium.

The choice rule of conditional cooperation I have posited assumes, in effect, the existence of a first-level (or "leader") metagame, which gives the follower a motive to cooperate against what Howard calls a "tit-for-tat" conditional strategy. But unlike Howard, I do not carry the analysis to a second-level (or "follower-leader") metagame in which the leader is given a motive to play tit-for-tat against the follower's own tit-for-tat policy, once removed.

The reason I eschew this stepwise backward reasoning is that it seems unnecessary if—as assumed of the Being in Newcomb's problem earlier—players' predictions (in the *preplay* leader-follower negotiation phase of the game) precede their choices (in the *play* of the game). Clearly, the proposal of conditional cooperation by the leader in the preplay phase is sufficient to initiate the process of cooperation. Then, however, the players become aware of each other's powers of prediction, prediction probabilities that satisfy the previous inequalities are sufficient to protect the players against either's reneging on an agreement that is reached. For given that each player knows that the other player's probability of predicting his own strategy choice is sufficiently high, he knows that he probably cannot "get away with" a sudden switch in his strategy choice in the play of the game, because this move already will have been anticipated with a high probability in the preplay phase. Hence, the assumption that (preplay) predictions precede (play-of-the-game) choices, and both players know this, deters "last-minute" intrigue that would render the cooperative outcome unstable.

The advantage offered by a leader-follower model that distinguishes unambiguously between the preplay and play phases of a game lies not only in its ability to truncate the iterative calculations of metagame theory. It also offers an advantage in highlighting the circumstances under which players would come to harbor tit-for-tat expectations in the first place. If they come to realize, in the preplay phase of the game, that their later choices in the play of the game are, to a sufficiently high degree, predictable, they will be robbed of their incentive to violate an agreement, given that they are expected-utility maximizers.

In this manner, the leader-follower model suggests circumstances under which an *absolutely* enforceable contract will be unnecessary. When the prediction probabilities of the players are sufficiently high (which depends on the utilities assigned by the players to the outcomes), an agreement to cooperate—reached in leader-follower negotiations in the preplay phase of the game—can be rendered "enforceable enough" so as to create a probabilistic kind of equilibrium that stabilizes the cooperative outcome.

By introducing probabilities of correct prediction *as parameters* in the preplay phase of a game, one is able to place the metagame solution to Prisoners' Dilemma within a rational-choice framework. What emerges as a solution is, in essence, a consequence of the rationality assumption (i.e., that players maximize expected utility) rather than the assumption that there exists some kind of consciousness of predictability among players. This is not to denigrate the metagame solution—which I regard as a major advance in game theory—but rather to show that there is a compelling rationale for its existence within the rational-choice framework.[5]

To what extent do players in real-world political games think in the terms I have described? This is a difficult question to answer generally, but one specific illustration of this kind of thinking may persuade the reader that it is certainly not unknown, at least in the field of foreign policy decision-making. In describing a highly classified mission, code-named Holystone, that allegedly involved reconnaissance by U.S. submarines inside Soviet waters, one U.S. government official was quoted as saying:

> One of the reasons we can have a SALT [Strategic Arms Limitation Talks] agreement is because we know of what the Soviets are doing, and Holystone is an important part of what we know about the Soviet submarine force [New York Times, May 25, 1975: 42].

In the final section, after first summarizing the preceding analysis, I shall touch on some implications of this remark.

## CONCLUSIONS

As originally formulated, Newcomb's problem suggests an apparent contradiction between the dominance principle and the expected-utility

[5] In fairness to Howard, he argues that metagame equilibrium choices are rational, but in a "stability" rather than an "expected-utility" sense (Howard, 1971: 61-63). The introduction of cardinal utilities (and probabilities), I believe, strengthens his rationality argument, though at the admitted cost of complicating his rather spare ordinal-game model.

principle. Following Ferejohn, however, I showed that the conflict between these two principles can be persuasively resolved if Newcomb's problem is reformulated as a decision-theoretic problem rather than as a game. In the decision-theoretic representation, which seems accurately to reflect the original statement of the problem, neither action is dominant for the person making the choice of which box(es) to choose. In the absence of a dominant strategy, therefore, the expected-utility principle cannot run amok of the dominance principle.

I then showed that the basic assumption about the predictability of choices in Newcomb's problem, when applied to not one but both players in Newcomb's prediction-choice game, defines a Prisoners' Dilemma in which the cooperative solution has considerable appeal. This appeal, to be sure, requires that one player (the leader) take the initiative and propose to the other player (the follower) a choice rule of conditional cooperation. It does not, however, require a binding and enforceable contract between the two players, which some analysts have argued is the only way to ensure cooperation. Nor does it require that the players rely solely on good will and mutual trust to bring about the cooperative outcome. Rather, the analysis suggests that there is a third (middle?) road to cooperation—mutual predictability of choices—that renders the cooperative strategies less risky for both players.

If such predictability obtains, then a contract is unnecessary, for violations will be predictable with a high probability in the preplay phase of the game and appropriate sanctions can be applied to the violator in the play of the game. But because such retribution works to the disadvantage of both players, the ability by both players to predict each other's choices serves also to reinforce trustworthy behavior, which is exactly what is not encouraged in Prisoners' Dilemma without mutual predictability.

I showed that this resolution of Prisoners' Dilemma bears some resemblance to the metagame solution to this game, but offers, in addition, a model that rationalizes its existence in terms of the expected-utility calculations of players. This has the advantage of placing the metagame solution within a probabilistic, rational-choice framework.

The kind of mutual predictability assumed in this leader-follower model, it seems, has given impetus to negotiations between the superpowers in SALT and laid the groundwork for certain arms-limitation agreements recently. With each superpower's reconnaissance satellites (and submarines, à la Holystone!) able to detect substantial violations quickly, the abrogation of an agreement by one party will be known before its consequences prove disastrous to the other party and prevent it from taking appropriate countermeasures. With little to be gained from such a violation and perhaps much to be lost, it is less likely to occur. In this manner, space-age technology has fostered arms-control agreements that—because of the ease with which violations could previously be kept secret—have been so difficult to obtain in the past.

Arms races are not the only situations that have the characteristics of a Prisoners' Dilemma game. Prisoners' Dilemmas have also been identified in diverse areas such as agriculture, business, law, and even the arts. Indeed, situations in which individuals can gain from cooperation, but also have an incentive not to cooperate to improve their payoffs still more, seem nearly universal.

Perhaps the most poignant statement of the problem inherent in such situations is the now famous article by Garrett Hardin (1968; see also Crowe, 1969). Although Hardin focuses on the population problem—contending that the social costs of overpopulation do not offer individual incentives for people to have fewer children—he treats this as one of a class of tragedy-of-commons problems "without a technical solution."

It would seem that the mutual-predictability model developed here does offer a "technical" solution, though for many-person games (like the population problem) it seems less meaningful and applicable. The reason is that leaders in such games are not so able as in two-person games to punish, by their own noncooperative actions, noncompliance by followers. Rather, it seems, followers as well as leaders would have to agree to the imposition of sanctions against noncompliance, enforceable by some higher authority (e.g., the state). Of course, sanctions that make it more rewarding to cooperate than not to cooperate—whatever the choices of the other players—transform Prisoners' Dilemma into another game.[6]

One may regard, of course, the mutual-predictability solution to Prisoners' Dilemma to be a paradox itself, because—as in Newcomb's problem treated as a game—it shows up a conflict between the dominance principle and the expected-utility principle. It seems more constructive, however, to stress the fact that one can calculate, from players' payoffs, probabilities that indicate thresholds at which the cooperative outcome in Prisoners' Dilemma can be rendered stable—and the paradox of noncooperation thereby circumvented. It is fitting, perhaps, that the predictability condition that is a central feature of one (apparent) paradox serendipitously offers, when applied to another paradox, the key that allows one to escape it.

[6] Relationships between two-person and n-person Prisoners' Dilemmas are discussed in Hamburger (1973) and Hardin (1971).

## REFERENCES

BAR-HILLEL, M. and A. MARGALIT (1972) "Newcomb's paradox revisited." British J. for the Philosophy of Sci. 23 (November): 295-304.

BRAMS, S. J. (1975a) "Deception in 2 x 2 games." New York University. (unpublished)

--- (1975b) Game Theory and Politics. New York: Free Press.

CROWE, B. L. (1969) "The tragedy of the commons revisited." Science 166 (November 28): 1103-1107.

GARDNER, M. (1974) "Mathematical games." Scientific Amer. 230 (March): 102-108.

--- (1973) "Mathematical games." Scientific Amer. 229 (July): 104-108.

HAMBURGER, H. (1973) "N-person prisoner's dilemma." J. of Mathematical Sociology 3: 27-48.

HARDIN, G. (1968) "The tragedy of the commons." Science 162 (December 13): 1243-1248.

HARDIN, R. (1971) "Collective action as an agreeable n-prisoner's dilemma." Behavioral Sci. 16 (September): 472-481.

HENDERSON, J. M. and R. E. QUANDT (1971) Microeconomic Theory: A Mathematical Approach. New York: McGraw-Hill.

HILL, W. W., Jr. (1975) "Prisoner's dilemma: a stochastic solution." Mathematics Magazine 48 (March): 103-105.

HOWARD, N. (1974) " 'General' metagames: an extension of the metagame concept," pp. 261-283 in A. Rapoport (ed.) Game Theory as a Theory of Conflict Resolution. Dordrecht, Netherlands: D. Reidel.

--- (1971) Paradoxes of Rationality. Theory of Metagames and Political Behavior. Cambridge: MIT Press.

LEVI, I. (1975) "Newcomb's many problems." Theory and Decision 6 (May): 161-175.

New York Times (1975) May 25: 1, 46.

NOZICK, R. (1969) "Newcomb's problem and two principles of choice," pp. 114-146 in N. Rescher (ed.) Essays in Honor of Carl G. Hempel. Dordrecht, Netherlands: D. Reidel.

RAPOPORT, A. (1974) "Prisoner's dilemma–reflections and observations," pp. 17-34 in A. Rapoport (ed.) Game Theory as a Theory of Conflict Resolution. Dordrecht, Netherlands: D. Reidel.

--- (1967) "Escape from paradox." Scientific Amer. 217 (July): 50-56.

--- and A. M. CHAMMAH (1965) Prisoner's Dilemma: A Study in Conflict and Cooperation. Ann Arbor: Univ. of Michigan Press.

SCHLESINGER, G. (1974) "The unpredictability of free choices." British J. for the Philosophy of Sci. 25 (September): 209-221.

SHUBIK, M. (1970) "Game theory, behavior, and the paradox of the prisoner's dilemma: three solutions." J. of Conflict Resolution 13 (June): 181-193.