1969

The Prisoner's Dilemma and Mutual Trust: Comment

Robert L. Birmingam
Indiana University School of Law

THE PRISONER'S DILEMMA AND MUTUAL TRUST: COMMENT

ROBERT L. BIRMINGHAM

IN "On the Meaning of Trust,"[1] Professor Held has recently defended the interpretation of the Prisoner's Dilemma game vigorously criticized by Professor Tullock in his comment[2] on three earlier articles by herself[3] and Professors Wolff[4] and Thompson.[5] Professor Tullock stated: "The error is a bit subtle. Each of the three articles begins with a perfectly correct account of the Prisoner's Dilemma. . . . [A]ll then make statements implying that the problem is one of mutual trust. This is simply not so."[6] Professor Held replies: "Along with others, I have described this problem in terms of whether or not it is rational to *trust*. . . . [I]t seems to me that trust is most required exactly when we least know whether a person will or will not do an action. And the Prisoner's Dilemma presents a paradigm of such a situation."[7]

In her response, Held relies unduly on a relatively fruitless definitional discussion of the term "trust." Her disagreement with Tullock appears more fundamental although in one sense equally semantic: the disputants are playing different games. The Prisoner's Dilemma arises through association of individual utility levels with alternative sentences. The basic matrix assigns years of imprisonment as a function of the choices confronting the parties involved. The formulation in Figure 1, for example, is appropriate when mutual silence and mutual confession respectively yield incarceration for one year and for six years to each criminal, while either combination of differing actions frees the confessing criminal and results in confinement of his partner for ten years. Here a and b designate the actions silence and confession open to players 1 and 2.

	a_2	b_2
a_1	(1, 1)	(10, 0)
b_1	(0, 10)	(6, 6)

FIG. 1

Tullock and Held would not dispute this initial characterization. Their differences are a product of attempts to translate years of imprisonment into units of utility. Tullock assumes that the value to each player of various outcomes is simply a decreasing function of his period of imprisonment. Application to the resulting schedule of the tautology that the individual will maximize his utility yields the traditional equilibrium of double confession. Tullock correctly concludes: "I may have the most perfect confidence that my fellow criminal will never confess without in any way affecting the desirability of my confessing. . . . In general . . . one prisoner's opinion about the probable behavior of the other is irrelevant to his own decision, since his payoff will always be higher if he confesses. . . . The problem raised by the dilemma is simply that if both parties make the same decision, they are better off if that double decision is 'don't squeal' than if it is 'squeal.' "[8]

The utilities derived by Held from the basic imprisonment pattern appear superficially consistent with those used by Tullock. The matrix she adopts, reproduced as Figure 2, preserves the classic solution. Moreover, she asserts: "Clearly, if one can make an accurate prediction either that he will or that he will not confess, one can decide in accordance with usual recommendations for rational behavior. . . . If one can accurately predict that one's fellow prisoner will confess, the rational course of action is also to confess, thus minimizing one's losses and avoiding the higher penalty of not confessing when he does. On the other hand, if one can accurately predict that he will *not* confess, the self-

	a_2	b_2
a_1	(5, 5)	(−10, 10)
b_1	(10, −10)	(− 5, − 5)

FIG. 2

ishly rational course of action is to confess."[9]

In this context other statements by Held seem incongruous:

> The interest of the Prisoner's Dilemma situation, however, is in considering what course of action may be deemed to be the rational one when one can *not* know what the other fellow will do. . . . [I]f the probabilities concerning the other fellow's behavior are either totally unknown or exactly .5, then the problem of establishing which course of action would be the rational one is acute. . . . When the chance of success is exactly even, and we are confronted with a one-shot decision, *should* we or *should* we not take a chance on furthering a common interest while risking an individual interest?[10]

One would suppose that if rationality compels confession when the other player is certain to confess or certain to remain silent, confession would similarly be indicated if his behavior is undetermined. This proposition can be simply demonstrated. Let prisoners 1 and 2 remain silent with probabilities x and y, respectively. The expected payoff to each can be written as a function of these probabilities:

$$E_1(x, y) = -5x + 15y - 5\,, \qquad (1)$$

$$E_2(x, y) = 15x - 5y - 5\,. \qquad (2)$$

Thus:

$$\frac{\partial E_1}{\partial x} = \frac{\partial E_2}{\partial y} = -5\,. \qquad (3)$$

Equation (3) demonstrates that E_1 varies inversely with x for all values of y, and E_2 varies inversely with y for all values of x. Therefore player 1 can maximize his expected payoff by setting x equal to 0, while player 2 can achieve his best position by choosing a similar value for y.[11]

Confusion arises because Held tacitly assumes utility to an individual to be a decreasing function of both his own sentence and the sentence of his partner. Thus she considers outcomes a_1b_2 and b_1a_2 not only damaging to the silent player but also less satisfactory than mutual silence to the confessing player. This assumption has been incorporated into the matrix of Figure 3.

	a_2	b_2
a_1	(5, 5)	(−10, 0)
b_1	(0, −10)	(−5, −5)

FIG. 3

The game has been transformed from the Prisoner's Dilemma into an approach to its solution. Mutual confession is now an obvious equilibrium position only in the sense that it will be reached if each player seeks to minimize his maximum loss. If again we posit that players 1 and 2 keep silent with probabilities x and y, we may write:

$$E_1(x, y) = 10xy - 5x + 5y - 5\,, \qquad (4)$$

$$E_2(x, y) = 10xy + 5x - 5y - 5\,. \qquad (5)$$

Hence:

$$\frac{\partial E_1}{\partial x} = 10y - 5\,, \qquad (6)$$

$$\frac{\partial E_2}{\partial y} = 10x - 5\,. \qquad (7)$$

Given the utility schedules of Figure 3, each player can expect to gain through silence so long as the probability that the other will not confess exceeds .5. Here suppositions of each prisoner concerning the choice to be made by his partner play a crucial role.

	a_2	b_2
a_1	(5, 5)	(5, −5)
b_1	(−5, 5)	(−5, −5)

FIG. 4

Further reduction of the value of unilateral betrayal or a lessening of its impact on the silent partner can yield an equilibrium of mutual silence. The impact of choice by an individual on his expected payoff in the game of Figure 4, for example, is independent of the probability of silence on the part of his partner. Thus:

$$E_1(x, y) = 10x - 5\,, \qquad (8)$$

$$E_2(x, y) = 10y - 5. \quad (9)$$

Consequently:

$$\frac{\partial E_1}{\partial x} = \frac{\partial E_2}{\partial y} = 10. \quad (10)$$

Each player will profit through selection of alternative a regardless of the behavior of the other.

Where interpersonal comparison of utilities is permissible, outcome a_1a_2 is typically taken to maximize the value of the Prisoner's Dilemma game to the players jointly. That their gain through movement from b_1b_2 to a_1a_2 may be more than offset by consequent loss to other members of society is indicated by the story illustrating the game itself and by Tullock's example of a competitive market. Where there are no counterbalancing external diseconomies, however, the community has an interest in converting the Prisoner's Dilemma game to a game such as that presented in Figure 4. One mechanism facilitating this transformation is the law of contract. If prisoners could bind themselves through an agreement not to confess, there would be no dilemma. The state, by requiring payment of damages for breach of certain promises, permits individuals to elect such an escape in many areas of interaction. Another solution, implicit in Held's analysis, would rely on socialization of the individual to induce incorporation of the welfare of others as an important element in his own preference function.

INDIANA UNIVERSITY SCHOOL OF LAW

NOTES

1. Virginia Held, "On the Meaning of Trust," *Ethics,* LXXVIII (January, 1968), 156.

2. Gordon Tullock, "The Prisoner's Dilemma and Mutual Trust," *Ethics,* LXXVII (April, 1967), 229.

3. Virginia Held, "Rationality and Social Value in Game Theoretical Analysis," *Ethics,* LXXVI (April, 1966), 215.

4. Robert Paul Wolff, "Reflections on Game Theory and the Nature of Value," *Ethics,* LXXII (April, 1962), 171.

5. George Thompson, "Game Theory and 'Social Value' States," *Ethics,* LXXV (October, 1964), 36.

6. Tullock, *op. cit.*, p. 229.

7. Held, "On the Meaning of Trust," p. 157 (emphasis in original).

8. Tullock, *op. cit.*, p. 229.

9. Held, "On the Meaning of Trust," p. 156 (emphasis in original).

10. *Ibid.*, pp. 156–57 (emphasis in original).

11. See Karl Henrik Borch, *The Economics of Uncertainty* (Princeton, N.J.: Princeton University Press, 1968), pp. 129–35.