Pure and Utilitarian Prisoner's Dilemmas

Steven T. Kuhn

Department of Philosophy

Georgetown University

Washington, D.C. 20057

Telephone: (202)687-7487

Serge Moresi

Department of Economics

Georgetown University

Washington, D.C. 20057

Telephone: (202)687-5830

Pure and Utilitarian Prisoner's Dilemmas

The prisoner's dilemma game has acquired large literatures in several disciplines.[1] It is a little surprising, therefore, that a good definition of the game is so hard to find. Typically an author relates a story about co-perpetrators of a crime or participants in an arms race, provides a particular payoff matrix and asserts that the prisoner's dilemma game is characterized by, or at least illustrated by, that matrix.[2] In the few cases in which characterizing conditions are given[3], the conditions, and the motivations for them, do not always agree with each other or with the paradigm examples in the literature. In this paper we characterize several varieties of prisoner's dilemma games. In particular, we suggest that there are at least two distinctions among prisoner's dilemma games that have

---

[1] (Donniger, 1986, p. 124) reports that "more than a thousand articles" were written on the subject in the 60's and 70's. Our own check under the keywords "prisoner's dilemma", "prisoners' dilemma" and "prisoner's dilemna" returned 116 entries in the *Social Science Index* since 1983, 118 entries in the *Econlit Index* since 1978, and 91 entries in the *Philosopher's Index* since 1982. This does not count material in the textbooks, dictionaries, encyclopedias, and handbooks of game theory, economics, psychology, philosophy and other disciplines.

[2]See, for example, (Gibbons, 1992, p. 3) and (Fudenberg and Tirole, pp. 9-10). The reader may also wish to consult these sources for standard terms and concepts of game theory that are used in this paper. The former is a very recent, carefully-written textbook emphasizing economic applications. The latter is a rigorous and comprehensive text and survey of the field. A thoughtful selection of other helpful sources is provided at the beginning of (Binmore, 1992).

[3] See (Rapoport and Guyer, 1966), (Rapoport, Guyer and Gordon, 1976), (Rapoport and Chammah, 1965) and (Axelrod, 1984). The first two consider only ordinal-valued payoffs and the other two consider only symmetric games. We allow cardinal-valued payoffs and asymmetric games. Other differences will be noted in the course of the paper.

philosophical significance, that between *pure* and *impure* dilemmas and that between *utilitarian* and *non-utilitarian* dilemmas. In the first section, we explain the distinctions and characterize them in terms of the payoff matrix. In the second, we discuss an issue of moral philosophy that illustrates the significance of the pure/impure distinction. In the third, we discuss an issue in law that illustrates the significance of the utilitarian/nonutilitarian distinction.

Taxonomy

The basic dilemma

Consider the following payoff matrix.

	C	D
C	R_r,R_c	S_r,T_c
D	T_r,S_c	P_r,P_c

Row and Column each choose whether to cooperate or defect. R_r and R_c are the *rewards* to Row and Column, respectively, for universal cooperation. P_r and P_c are their *punishments* for universal defection. T_r and T_c are the *temptation* payoffs that Row and Column receive for defecting while the other cooperates. S_r and S_c are the *sucker* payoffs they get for cooperating while the other defects.

A minimal set of conditions needed for this game to be considered a prisoner's dilemma might require that defection *dominate* cooperation for each player, but that universal cooperation be unanimously preferred to universal defection. In other words:

B1. $T_r>R_r$ and $P_r>S_r$

B2. $T_c>R_c$ and $P_c>S_c$

B3. $R_r>P_r$ and $R_c>P_c$.

We label a game meeting conditions B1-B3 a basic prisoner's dilemma.[4] The conditions for a game to be a basic prisoner's dilemma do not require either cardinal or interpersonal utility measurements. They require only that each player rank its own temptation, reward, punishment, and sucker payoffs in descending order. In games meeting these conditions, defection is the dominant choice for each player: each is better off defecting, whether the other cooperates or defects.[5]

---

[4]In each of conditions B1,B2 and B3, one of the two strict inequalities could be replaced by a weak inequality without losing the character of the dilemma. Games satisfying B1-B3 might thus be said to be *strict* prisoner's dilemmas: if the payoffs are changed by a sufficiently small value, the game will continue to satisfy the conditions.

[5]The argument that rational players defect in a basic prisoner's dilemma does not depend on the either player's knowing the other is rational. If we add a weak "common knowledge" assumption--that each knows the other is rational--conditions B1-B3 can be significantly weakened. Let an *extended* prisoner's dilemma be a game that satisfies:

E1. $T_i>R_i$ and $P_i>S_i$ for some i=r,c

E2. $P_j>S_j$ for $j\neq i$

E3. $R_r>P_r$ and $R_c>P_c$

Since j knows that i is rational, E1 implies that j will expect i to defect. In that case, E2 implies that j will defect himself and E3 implies each player will be worse off than they would have been had they cooperated. (Again, a "weak" version of the conditions could be obtained by allowing one of the strict inequalities in E1 and E2 to be replaced by a weak inequality.) In the exhaustive classifications of 2x2 games of (Rapoport, et al., 1966) and (Rapoport, et al., 1976), these games are not classified as prisoner's dilemmas because one player could induce the other to cooperate by "threatening" to cooperate herself. But if the players are known to be rational, such a threat

Universal defection is the only Nash equilibrium, i.e., the only outcome in which neither player has reason to regret its choice (given the choice of the other).[6] However, universal cooperation is unanimously preferred to mutual defection, which implies that the equilibrium outcome is not (pareto) efficient. Under these conditions one feels the force of the dilemma: players who *aim* at self-benefit choose defection, but they would *achieve* more self-benefit if they each choose cooperation. It is often implicitly assumed that universal cooperation is the most desirable outcome of a prisoner's dilemma game. On any reasonable definition, a most desirable outcome will be Pareto efficient. As will be shown in the next section, however, the definition of the basic game (B1-B3) does not imply that universal cooperation has this property.

The pure dilemma

There is no reason why rational players in a prisoner's dilemma need restrict themselves to either certain cooperation or certain defection. Each of them might consider a *mixed* strategy of cooperating with probability p and defecting with probability 1-p. When such independent mixed strategies are permitted, the argument that B1-B3 ensure that universal cooperation is the optimal outcome no longer goes through. Let us say that a basic prisoner's dilemma is *pure* if there is no pair of independent mixed strategies that provides both players higher payoffs than they get from universal certain cooperation, i.e., when universal cooperation is Pareto efficient relative to the class

---

would not be credible.

For convenience, we restrict our attention in this paper to basic prisoner's dilemma games. We believe, however, that the discussion could be generalized to include all extended prisoner's dilemmas.

[6]And in fact, universal defection is a *dominant equilibrium* (or equilibrium in dominant strategies) which is a stronger equilibrium concept than Nash equilibrium.

of mixed strategies.[7] In this section we demonstrate that not all prisoner's dilemmas are pure and we characterize pure dilemmas in terms of their payoff matrices.

The payoffs of a basic prisoner's dilemma are shown in figure 1. Row's utility is plotted along the horizontal axis and Column's is plotted along the vertical axis. The labeled points correspond to the four payoffs that are possible when mixed strategies are not permitted. For example, the point labeled (C,D) represents the payoff (S_r,T_c) that results from Row's cooperating and Column's defecting. Condition B1 guarantees that (C,D) lies to the left of (D,D); B2 and B3 guarantee that it lies above. Similarly, B2 guarantees that (D,C) lies below (D,D); B1 and B3 guarantee that it lies to the right. B3 guarantees that (C,C) lies above and to the right of (D,D). B1 and B2 guarantee that it lies beneath (C,D) and to the left of (D,C). The solid lines connecting the labelled points represent the feasible payoffs when *one* of the players adopts a mixed strategy. For example the points on the line from (D,D) to (D,C) at the bottom of the figure represent the feasible payoffs when Row defects and Column mixes. The points along the line from (C,D) to (C,C) represent the payoffs when Row cooperates and Column mixes. The payoff when Row defects and Column cooperates with probability ¼ and the one when Row cooperates and Column cooperates with probability ¼ are marked with x's. The line connecting these two points represents the set of feasible payoffs when Column cooperates with probability ¼ and Row mixes. Similarly, the line connecting the two points marked with an asterisk represents the set of feasible payoffs when Column cooperates with probability 3/4 and Row mixes. Notice that the former line lies entirely within the odd-shaped quadrilateral formed by the original four points, whereas the latter line does not. Indeed, the latter line contain points that lie northeast of (C,C). Thus if both players adopt mixed strategies they can achieve higher expected utilities than under universal cooperation. The prisoner's dilemma depicted is therefore not pure: we call it *impure*.

---

[7]We make the standard assumption that preferences over uncertain outcomes are represented by the expected utility of these outcomes. When no confusion will result, we simply write "utility" or "payoff" for "expected utility" or "expected payoff".

The payoffs that are feasible when arbitrary independent mixed strategies are permitted are those that lie in the region bounded by the four solid line segments and the connecting curve in figure 2. When (C,C) lies outside the triangle formed by (D,C), (D,D) and (C,D) (i.e., northeast of the uppermost dotted line in the figure) the feasible payoffs under independent mixed strategies are within the quadrilateral bounded by those four points. When (C,C) lies within that triangle, however, the feasible payoffs extend beyond (C,C). Thus a prisoner's dilemma game is pure if and only if (C,C) lies outside the triangle.

The question of whether the point lies outside or inside the triangle is equivalent to the question of whether the line from (C,D) to (C,C) has a shallower slope than the line from (C,C) to (D,C) or a steeper one. So the game is pure if and only if[8]:

$$P) \quad \frac{T_c - R_c}{R_r - S_r} < \frac{R_c - S_c}{T_r - R_r}$$

T_i-R_i can be thought of as the *temptation bonus* to player i. It is the amount of additional utility (relative to the reward) he gains from defecting. Similarly R_i-S_i is the *sucker penalty*. It is the amount i loses (relative to the reward) by the other's defection. Using T^+ and S^- for the temptation bonus and sucker penalty, and rearranging terms, we see that the condition for purity can be expressed:

$$P') \quad S_r^- S_c^- > T_r^+ T_c^+,$$

or, equivalently, that the (geometric) mean sucker penalty exceeds the (geometric) mean temptation bonus. An alternative demonstration that P' characterizes purity, which exploits an analogy between the prisoner's dilemma game and a pure exchange economy, is presented in an appendix.

Note that condition P, unlike B1-B3, requires cardinal utilities. It does not require that the utilities

[8]We use strict inequalities in P and P' below to facilitate subsequent comparison with RCA, although the gloss above in terms of Pareto efficiency would suggest weak inequalities. Nothing substantive hinges on the choice.

be interpersonally comparable, however. Changing the units in which r's utility is measured (i.e., subjecting them to a linear transformation) would not affect the inequality in P.

In a basic prisoner's dilemma, no pair of pure strategies benefits both players more than cooperation by both. In a pure prisoner's dilemma, no pair of mixed strategies benefits both players more than certain cooperation by both. Under some conditions players may be able to use *correlated* mixed strategies. For example, if players r and c are able to condition their strategies on a publicly observable event, they could achieve outcome (C,D) with probability p and outcome (D,C) with probability 1-p. One might think that such conditions would require a third characterization of the notion of prisoner's dilemma, i.e., that the conditions under which no mixed correlated strategy is better for each player than universal cooperation might be different than the conditions under which no pair of independent mixed strategies is better. This impression, however, is mistaken. If mixed correlated strategies are allowed in the game of figure 3 the set of possible payoffs remains exactly the same. If they are allowed in the game of figure 2, the set of possible payoffs corresponds to the region within the triangle (D,D)-(D,C)-(C,D). So again, universal cooperation is optimal if and only if (C,C) lies outside the triangle. Correlated mixed strategies may (as they do in figure 2) make available payoffs superior to the payoffs from universal cooperation and to any payoffs that can be obtained from independent mixed strategies. Whenever a correlated strategy benefits both parties more than universal cooperation, however, there are independent strategies that do so as well.

Correlation is particularly important for repeated games, where players know that later rounds can be used to reward and punish the play of current rounds. For example, in a finitely repeated prisoner's dilemma two players could, by taking turns as defector and cooperator, assure themselves payoffs $S_r+T_r+...+S_r+T_r$ and $S_c+T_c+...+S_c+T_c$. These considerations have led a number of authors[9]

[9] Precisely this argument is given in (Rapoport et al., 1965, pp. 34-35) and (Axelrod, 1984, p. 10). (This is why we label the condition RCA.) In the former work, the condition is viewed as "an additional constraint" whereas in the latter it is viewed as part of the definition of the

considering symmetric prisoner's dilemmas (those in which $S_r=S_c=S$, $T_r=T_c=T$, and so on) to require the condition:

RCA) $R>\frac{1}{2}(S+T)$.

This ensures that the correlating turn-takers will do worse than cooperators. But the arguments above show that correlation can beat universal cooperation if and only if independent strategies can, and that this happens if and only if condition P is not met. So one need not have used correlation to motivate condition RCA. Indeed, a quick check reveals that, in the symmetric case, P *is* RCA.

The utilitarian dilemma

Suppose we are strict Benthamite utilitarians, i.e., we believe that the most desirable outcomes are those that maximize total utility, regardless of how that utility is distributed. Then the observation that universal cooperation is efficient with respect to pure strategies or mixed strategies, would not be enough to ensure that it is a most desirable outcome. More total utility might be

---

prisoner's dilemma game. Some similar conditions that look deceptively like appropriate generalizations of RCA to the nonsymmetric case also appear in the literature. Let RCA^* be the condition $R_i>\frac{1}{2}(T_i+S_i)$ for i =r,c. As noted in [], RCA^* can be extracted from the remark in (Gauthier, 1967) that each player loses less from his own cooperation than he gains from the other's cooperation. Let RCA^{**} be the condition $R_i<\frac{1}{2}(T_i+S_i)$ for i =r,c. Sobel (1991, p. 34) labels a basic prisoner's dilemma that meets RCA** a *stretched* prisoner's dilemma and notes that in a stretched prisoner's dilemma both players do better from a fifty-fifty mix of the correlated strategies (D,C) and (C,D) than they do from universal cooperation. (In fact RCA^{**} is both necessary and sufficient for the superiority of the fifty-fifty mix.) He also notes that in stretched dilemmas meeting the requirement $R_i<\frac{1}{4}(T_i+S_i+R_i+P_i)$ for all i=r,c, both players do better adopting a fifty-fifty mix of the independent strategies C and D than they do cooperating. As noted in this paper, mixed strategies (whether independent or correlated) can surpass universal cooperation if and only if a prisoner's dilemma is impure.

associated with outcomes (C,D) or (D,C) (or both) than with (C,C). (And (C,D) and (D,C) are also Pareto efficient). An example of such a game is shown in figure 3. Here the total utility at (D,C) is S_r+T_c. The higher of the two dotted lines contains all the points at which the sum Row's and Column's utility would have this value. Since (C,C) lies southwest of this line, it represents a smaller total utility and, from the utilitarian perspective, a less desirable outcome.

Accordingly, let us define a *utilitarian* prisoner's dilemma as a game in which, in addition to conditions B1-B3, (R_r,R_c) has the greatest total utility of all possible outcomes. Since B3 implies $P_r+P_c < R_r+R_c$, and since any mix of T_r+S_c and S_r+T_c lies between these two values, the utilitarian dilemma can be characterized by the conditions B1-B3 and:

U) $(R_r+R_c)>(T_r+S_c)$ and $(R_r+R_c)>(S_r+T_c)$.

In figure 3, S_r+T_c is larger than T_r+S_c. For (R_r,R_c) to maximize utility, (C,C) must lie northeast of the higher dotted line (say at point Q), and so the line from (C,D) to (C,C) must have a slope shallower than 1 and the line from (C,C) to (D,C) must have slope steeper than 1. Hence $\frac{T_c-R_c}{R_r-S_r}<1$ and $\frac{R_c-S_c}{T_r-R_r}>1$. Rearranging terms gives us the two clauses of condition U. A different rearrangement yields $T_r^+<S_c^-$ and $T_c^+<S_r^-$, i.e., the temptation bonus for each player is less than the sucker penalty for the other. Note that these conditions, unlike B1-B3 and P, do require that utilities be interpersonally comparable.

If a game fails to satisfy condition U, strict utilitarians would not regard it as true prisoner's dilemma since they prefer the outcome in which one party defects (with probability one) and the other cooperates to the outcome in which both cooperate. A non-utilitarian may not be impressed by this observation. The outcomes (C,D) and (D,C) represent the most unequal distributions of utility, so they do not seem particularly desirable. If utilities are costlessly transferrable, however, the distinction between utilitarian and non-utilitarian prisoner's dilemmas is significant for the non-

utilitarian as well. For under these conditions the issues of efficiency and equity can be separated. The most desirable solution to the *initial* game is the one that maximizes utility In figure 3 this is outcome (D,C). After the initial game the players can redistribute utility, moving northwest along the upper dotted line until their equity concerns, whatever they might be, are satisfied. As long as the game is not utilitarian, they could reach a point northeast of (C,C), i.e., an outcome they both prefer to universal cooperation. Thus there can be two motives for focussing on utilitarian games. For utilitarians, they are games in which universal cooperation is the most desirable outcome, and for non-utilitarians they are games in which universal cooperation is the most desirable outcome even when utilities are transferrable.

Each kind of prisoner's dilemma we have discussed can be obtained by adjusting the payoffs for universal cooperation. Figure 3 represents a dilemma that is pure, but non-utilitarian. If (C,C) were below the lower dotted line, say at point P, the resulting game would be an impure and non-utilitarian dilemma. If it were above the upper dotted line, say at point Q, the game would be a pure and utilitarian dilemma. In the symmetric case condition U, like condition P, reduces to condition RCA, and so for symmetric games the notions of pure and utilitarian dilemma coincide. In general, however, as figure 3 shows, the class of pure dilemmas properly contains the utilitarian ones. The relations among the various games is summarized in figure 4.

Morality and the prisoner's dilemma

The prisoner's dilemma is often invoked, particularly by defenders of social contract views of morality[10], to show the sense in which moral institutions are mutually advantageous. In Book II of

[10]Perhaps the best known invoker is David Gauthier. In (Gauthier, 1967) the game is used to explicate ideas attributed (approvingly) to Kurt Baier. It is not clear whether the prisoner's dilemma really does provide the appropriate model for morality in Gauthier's later writings because he now maintains that a person's disposition to cooperate with similarly disposed people

The Republic, Glaucon describes to Socrates "what people consider the origin and nature of justice."

> "They say that to do wrong is naturally good, to be wronged is bad, but the suffering of injury so far exceeds in badness the good of inflicting it that when men have done wrong to each other and suffered it, and have had a taste of both, those who are unable to avoid the latter and practice the former decide that it is profitable to come to an agreement with each other neither to inflict injury nor to suffer it...This, they say, is the origin and essence of justice; it stands between the best and the worst, the best being to do wrong without paying the penalty and the worst to be wronged without the power of revenge. The just then is a mean between two extremes;.."[11]

It does not seem unreasonable here to identify "best," "worst" and "mean" with the temptation, sucker and reward payoffs of a prisoner's dilemma.

Of course universal cooperation is not regarded as the outcome dictated by morality in every situation that can be modeled by a prisoner's dilemma game. In the formulation that gives the game its name, the "cooperative" solution is the one in which two criminals each refuse to confess to a murder they have jointly committed. In examples from economic literature, it is one in which two producers stick to fixed prices. Morality does not seem to dictate "cooperation" in either case. But in these examples the players' actions affect others. In the artificial world in which the only individuals, or the only ones who matter, are the players, refusing to confess or to lower prices seems the right thing to do. In the real world it may be possible to show that, in some larger game, confessing wrongdoing and refusing to fix prices are themselves the cooperative choices.

The arguments that choices between moral and amoral behavior resemble choices between

---

is, to some degree, detectable by others. The game is also frequently invoked in explications of Thomas Hobbes. See, for example, (Kavka, 1986).

[11](Plato, 358e-359b).

cooperation and defection in prisoner's dilemma games, however, do not generally show that the games in question are pure dilemmas. It is easy to maintain that a person would benefit greatly if he alone broke an agreement, somewhat less if both parties kept it, still less if neither did, and that he would suffer the most if he alone kept it. It is not so easy to maintain that the payoffs satisfy condition P or P'.[12] If moral choice does sometimes resemble play in impure prisoner's dilemmas, then we know that both parties will benefit more from randomized strategies than they do from pure ones. There are two possibilities. Either morality calls for randomized behavior or there is a gap between morality and advantage. Either possibility would have implications for moral theorizing. The former would undercut the notion that morally correct behavior is best described by systems of "deterministic" commands or principles. Ethical theories may order us to *keep your promises*; *avoid injury to others*, *act in the way you wish others to act* or they may inform us that *it is a duty to reciprocate kindness*; *it is wrong to ignore suffering*. They do not, however, order us to consult a randomizing device before deciding whether to keep a promise or inform us that it is right to reciprocate kindness 90% of the time. The latter would undercut the notion that morality is somehow based on (or even consistent with) rationality. How can it be rational to adopt a system that calls for one pattern of behavior when a different pattern of behavior is unanimously preferred?

---

[12]It might be possible to argue that moral considerations arise in situations meeting some condition simpler than P that nevertheless implies P. Indeed, in the quoted passage from Plato, one might take the phrases "suffering of injury" and "good of inflicting it" to be referring to values of S and T *relative to the reward payoffs.* In that case the first sentence might be interpreted as saying that, when S^+ is much larger than T^- for all, then experienced men of limited power will agree to cooperate. The condition that $S_i^+ > T_i^-$ for all i obviously implies P'. In fact it is equivalent to condition RCA^* of footnote 8. It is implausible to suppose that morality would endorse the cooperative action *only* when such a strong condition obtains. If it did, however, then our discussion of purity would show that in some instances everyone would benefit if everyone performed actions that morality did *not* endorse.

Legal contracts and the prisoner's dilemma

Consider two firms, Row and Column, engaged in a joint venture. The firms are pooling resources and working on a common research project. The expected outcome is a new line of products and higher profits for both firms. Given the current state of the research, however, each firm faces a dilemma: it can secretly start its own project, undermining the joint venture. We say that a firm defects (D) if it starts its own project, and cooperates (C) if it does not. For simplicity, we assume the following (expected) payoffs:

	C	D
C	2,2	$0,T_c$
D	T_r,0	1,1

where T_r and T_c are greater than 2 so that we have a prisoner's dilemma.[13]

The dilemma is pure if and only if $(T_r-2)(T_c-2)\leq 4$. It is utilitarian if and only if $\max(T_r,T_c)\leq 4$. Since each choice of T_r and T_c corresponds to a different prisoner's dilemma, our taxonomy can be illustrated as in figure 5. Values of T_r are plotted on the horizontal axis and values of T_c on the vertical. Each point represents a prisoner's dilemma game. The curved line is the set of games for which $(T_r-2)(T_c-2)=4$, so the impure dilemmas are the points northeast of the curve. The utilitarian dilemmas, for which T_r and T_c are both less than 4 occupy the square in the southwest. We consider

[13]We have in mind ventures like the one undertaken in 1981 between IBM and Microsoft Corporation that led to the joint development of the DOS operating system, and the one undertaken by the same firms in 1987 that was supposed to lead to the joint development of OS/2.

four particular prisoner's dilemmas, corresponding to the points P_1, P_2, P_3, and P_4 (so that each temptation payoff can be either 3 or 5). Note that only P_4 is impure, only P_1 is utilitarian and P_2 and P_3 are pure but not utilitarian.

In the absence of a binding contract, rational firms would defect. We thus assume that a contract can be written and, if one firm defects, the other firm can prove it in court. A *legal system* can be viewed as an institution that enforces monetary transfers between the parties. In our example, a *law* or *contract* determines the amount t that a defecting firm must pay to a cooperating firm.[14] In other words, a legal system transforms the original dilemma into the following game.[15].

	C	D
C	2,2	$t, T_c - t$
D	$T_r - t, t$	1,1

For j=1,...,4 we let G_j be the transformation of the game P_j by payment t. If t is relatively small--in our example, if t<1--then G_j is still a prisoner's dilemma. Defection is still the dominant strategy so that the law should have no effect on behavior. Each firm obtains a payoff of 1 as it would in the

[14]In principle the firms could write a *complete contract*, that is, a contract specifying monetary transfers between the parties conditional on their behavior and on the realization of any T_r and T_c. Although the values of T_r and T_c (like all the payoffs) are assumed to be known to the firms, it may be difficult or impossible to verify them in court, making complete contracts difficult to enforce. For that reason we restrict attention here to contracts with fixed t. (In a general setting, we would expect the value of t to vary with the entire payoff matrix. For simplicity, we assume that only temptation payoffs are unavailable to the courts.)

[15]For simplicity, legal fees, administrative costs, and penalties paid to the state or to third parties are assumed to be negligible relative to the transfer t.

absence of a legal system. However, if t is sufficiently large (t>3), cooperation becomes the dominant strategy. The law affects behavior and each firm's payoff is 2. Such "harsh punishments" are optimal when the dilemma is known to be utilitarian, as P_1. In this case the law has a *preventive* role and the court is ideally never asked to enforce the law. If the dilemma is not utilitarian (as P_2, P_3 and P_4) harsh punishments are not optimal. In G_2, each firm obtains a payoff of 2 if t>3 (see above) while it obtains a payoff of 2.5 if t=2.5. (Cooperation is a dominant strategy for Column and Row's best response is to defect.)[16] A similar argument applies to game G_3. In game G_4, if t=2.5 then both (D,C) and (C,D) are Nash equilibria. This means that the law may transform a prisoner's dilemma into a "coordination game"[17]. Assuming that the firms can coordinate their behavior on one of these two equilibria, they both obtain a payoff of 2.5, while they would obtain only 2 if t>3.

The above examples show that the distinction between utilitarian and non-utilitarian dilemmas is important in determining the role of the law. Utilitarian dilemmas call for a preventive role whereas non-utilitarian dilemmas call for a compensatory role. This does not necessarily mean that utilitarian and non-utilitarian dilemmas require different contracts. Suppose, for example, that Row and Column know that they will face a dilemma that will be P_1 with probability p (0<p<1) and either

---

[16]If the objective is simply to maximize the total payoff any t between 1 and 3 is equally acceptable. One firm obtains a payoff of 5-t and the other firm obtains t. For 2<t<3 both firms obtain more than under universal cooperation. For t=2.5, one firm will defect and the total gain is split equally between the two firms.

[17]Indeed, such a transformation is unavoidable in the sense that (D,C) will be a Nash equilibrium if and only if 1<t<3, and this will obtain if and only if (C,D) is also a Nash equilibrium. The reader should not assume that this phenomenon is tied to the pure/impure distinction. Some impure games are transformed into games in which, say, (C,D) is the only Nash equilibrium.

P_2, P_3 or P_4 with probability 1-p. Now suppose they set t=2.5. Then if the actual game turns out to be G_1 both players will cooperate and receive 2 units; if it is G_2, Row will defect and both will receive 2.5; if it is G_3, Column will defect and both will receive 2.5; and if it is G_4, one or the other will defect and both will receive 2.5. Thus, the *ex ante* expected payoff of each firm is given by 2p+2.5(1-p), which is greater than 2. This is an optimal contract. If t>3, both firms will always cooperate and the *ex ante* expected payoff for each is equal to 2, and if t<1 both will defect and their expected payoff is equal to 1. Intuitively, t=2.5 is sufficiently high to prevent any defection when the dilemma is utilitarian and, at the same time, it is sufficiently low to induce one firm to defect when the dilemma is not utilitarian.

Summary and conclusion

The prisoner's dilemma game is commonly used to illustrate the divergence between individual and collective rationality. The various examples found in the literature have a common feature: individual rationality calls for defection (regardless of the other player's strategy) and, if both players defect, both are worse off than if they had cooperated. We have taken this condition to characterize a *basic* prisoner's dilemma.

In a basic prisoner's dilemma, however, universal cooperation (with probability one) may itself violate collective rationality. That is, there may be some *mixed* strategies such that, if both players adopt them, both are better off (ex ante) than if they had cooperated with certainty. When no such mixed strategies exist, we call the prisoner's dilemma a *pure* dilemma. When such strategies do exist, we call it an *impure* dilemma.

It turns out that in a pure dilemma, *if utilities are transferable*, universal cooperation may still violate collective rationality. That is, if one player defects while the other cooperates, there may be some transfer from the defector to the cooperator (e.g., some monetary payment) such that, if the transfer takes place, both players are better off than if they had cooperated. When no such transfer exists, we call the prisoner's dilemma a *utilitarian* dilemma. When such a transfer does exist, we

call it a *nonutilitarian* dilemma.

We believe that the above distinctions are relevant in many applications of the prisoner's dilemma game. For example, if a moral rule prescribes behavior in impure dilemma situations, then *either* moral rules call for randomized behavior *or* there is a gap between morality and mutual advantage. To take another example, when the parties to a legal contract face a nonutilitarian dilemma, low damage awards may benefit both parties more than high damage awards. Both of these examples, of course, could be given more detailed analyses than we have done here. We wanted only to provide a precise definition of the prisoner's dilemma game, to show that there are interesting structural differences among prisoner's dilemmas and to show that it may be illuminating to pay attention to these differences when the game is employed in various disciplines.

Appendix: Proof that condition P' characterizes the PD's for which (C,C) is pareto efficient relative to both the mixed and correlated strategy sets.

Since the set of correlated strategies contains the set of mixed strategies we need only show that (a) P' is necessary in the mixed case and (b) P' is sufficient in the correlated case.

(a) Let us represent a mixed strategy for player i (i=r,c) by the probability $q_i \in [0,1]$ that she plays D. For any strategy profile $(q_r,q_c) \in [0,1]^2$, let us denote the players' expected payoffs by $U^r(q_r,1-q_c)$ and $U^c(1-q_r,q_c)$. This notation shows that the set of feasible payoffs is identical to the set of feasible utilities in a fictitious pure exchange economy. This economy is endowed with 1 unit of "good r" and 1 unit of "good c", and there are two consumers with preferences given by the utility functions U^r and U^c. Furthermore,

$$U^r(q_r,1-q_c) = (1-q_r)(1-q_c)R_r + (1-q_r)q_cS_r + q_r(1-q_c)T_r + q_rq_cP_r, \quad (1)$$

$$U^c(1-q_r,q_c) = (1-q_r)(1-q_c)R_c + (1-q_r)q_cT_c + q_r(1-q_c)S_c + q_rq_cP_c, \quad (2)$$

and it is easy to check that U^r and U^c are continuous and strictly monotonic (recall that $T_i>R_i>P_i>S_i$). Hence a PD is pure only if $MRS^r(0,1) \leq MRS^c(1,0)$, where MRS^i is consumer i's marginal rate of substitution in the fictitious economy. Using (1) and (2), this inequality implies condition P'.

(b) Suppose that unversal cooperation is *not* pareto efficient, i.e., there exists a probability distribution $(q_{CD},q_{DC},q_{CC},q_{DD})$ over the pure-strategy outcomes ((C,D)(D,C),(C,C),(D,D)) which both players prefer to (C,C). Since (D,D) is pareto dominated by (C,C), there must be such a probability distribution in which $q_{DD}=0$. Hence, there exist $q_{CD} \geq 0$, $q_{DC} \geq 0$ and $q_{CC}=1-q_{CD}-q_{DC} \geq 0$ such that

$q_{CD}S_r + q_{DC}T_r + (1-q_{CD}-q_{DC})R_r > R_r$, and

$q_{CD}T_c + q_{DC}S_c + (1-q_{CD}-q_{DC})R_c > R_c$.

These inequalities contradict condition P'.

Axelrod, Robert 1984. *The Evolution of Cooperation*. New York: Basic Books.

Binmore, Kenneth 1992. *Fun and Games*. Boston: D.C. Heath and Company.

Donniger, Christian. 1986. "Is it Always Efficient to be Nice? A Computer Simulation of Axelrod's Computer Tournament." In Paradoxical Effects of Social Behavior: Essays in Honor of Anatol Rapoport, edited by Andreas Diekmann and Peter Mitter, pp187-198. Vienna: Physica-Verlag.

Fudenberg, Drew and Tirole, Jean 1991. *Game Theory.* Cambridge, Mass.: The MIT Press.

Gauthier, David 1967. "Morality and Advantage." *Philosophical Review* 76:460-475.

Gibbons, Robert 1992. *Game Theory for Applied Economists*. Princeton, N.J.: Princeton University Press.

Kavka, Gregory 1986. *Hobbesian Moral and Political Theory*. Princeton, N.J.: Princeton University Press.

Plato. *The Republic*. G.M.A. Grube, translator. Indianapolis: Hackett Publishing Company.

Rapoport, Anatol and Chammah, A.M. 1965. *Prisoner's Dilemma.* Ann Arbor: University of Michigan Press.

Rapoport, Anatol and Guyer, Melvin J. 1966. "A Taxonomy of 2X2 Games." *General Systems* 11:203-14.

Rapoport, Anatol, Guyer, Melvin J., and Gordon, David G. 1976. *The 2X2 Game.* Ann Arbor: The University of Michigan Press.

Sobel, J. Howard , 1991. "Constrained Maximization." *Canadian Journal of Philosophy* 21: 25-52.