The Impure Game: Feasible Payoffs and Possible Generalizations [*]

Steven T. Kuhn

August 25, 2005

Abstract

A prisoner's dilemma game is said to be impure if there are independent mixed strategies that provide both players a higher payoff than does mutual cooperation. The phenomenon of impurity has important implications for the application of game theory to normative disciplines. This paper outlines a calculation (best performed by machine) that provides a full (if unwieldy) characterization of the space of outcomes in any impure prisoner's dilemma and applies this result to a simple example. Outcomes with particularly desirable properties are identified and described. The paper concludes with an examination of the extent to which the idea of impurity can be carried over to games other than the prisoner's dilemma.

1 Introduction: Pure and Impure Games

In [4], [3] and [2] a prisoner's dilemma is labelled *impure* if there are independent mixed strategies that are pareto superior to mutual cooperation. For example, figure 1 below gives the payoff matrix for a prisoner's dilemma game when the sucker, punishment, reward, and temptation payoffs for each player are 0, 1, 3 and 7. A simple calculation reveals that, if each player employs a strategy of 90% cooperation then each gets an expected payoff of 3.07, which is slightly better than the three units each gets by certain cooperation.

	C	D
C	3,3	0,7
D	7,0	1,1

Figure 1: An Impure Prisoner's Dilemma

This phenomenon would seem to be of great significance for the often expressed hope that game theory might provide a rigorous foundation (or at least

[*] Thanks to Donald Saari for insightful conversations and suggestions that made this paper possible.

an important tool) for moral philosophy. Conflict and congruence of interest and advantage is, after all, a central concern of moral philosophy and game theory is the theoretical discipline that aims to study conflict and congruence of interest. The prisoner's dilemma, in particular, has been said to illustrate how moral rules allow individuals to achieve advantages that selfish behavior precludes, and thus to suggest to suggest the senses in which the old question "why should I be moral?" can and cannot be answered. [1] Viewed in this light, the impure prisoner's dilemma shows that, unless moral rules sometimes require randomized choice, individuals who do not consistently follow such rules can achieve even greater mutual advantage than those who do consistently follow them. Thus the pure/impure distinction would seem to play an important role in applications of game theory to moral philosophy and impure games would seem to provide an key test for such applications.

From this perspective, several questions about the pure/impure distinction are of particular interest. One wants to know, first of all, the conditions under which impure versions of the prisoner's dilemma occur. This question is answered in [4], where it is established that a prisoner's dilemma is impure if and only if the following simple condition condition is satisfied.

$$I) \quad (T_x - R_x)(T_y - R_y) > (R_x - S_x)(R_y - S_y)$$

where S_i,R_i, and T_i are the sucker, reward and temptation payoffs for player $i = x, y$.

One also wants to know generally what payoffs are achievable in an impure dilemma and what outcomes might be considered desirable "solutions" to the games in view of these payoffs. These latter questions are not resolved in the earlier literature and they turn out to have to have answers that are much less simple. Finally, one wants to know the extent to which the idea of impurity is a general concept that extends to games other than the prisoner's dilemma. This paper continues the investigation of impure games begun in [4], [3], and [2] by addressing these central questions. Section two provides a full characterization of the space of payoffs attainable by independent mixing in the impure prisoner's dilemma. The characterizing equation is not easily readable, but its general form is discernable and it yields simple formulas for particular games. Section three briefly examines some plausible solutions for impure prisoner's dilemmas. Section four begins an investigation of the extent to which the idea of impurity extends to games other than the prisoner's dilemma.

2 Characterization of Payoff Space

The space we wish to characterize is illustrated in figure 1 below. The x and y axes represent the payoffs to players x and y. The dark lines trace a concave quadrilateral. The four vertices are the payoffs to the pure strategy outcomes in an impure prisoner's dilemma, i.e., the payoffs to (C, D), (C, C), (D, C), and

[1] See, for example,[1].

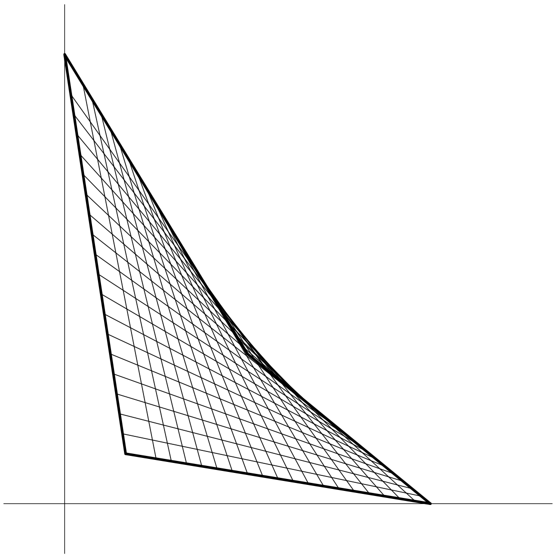

Figure 2: Feasible Payoffs for the Impure PD

(D, D), clockwise beginning at the top left. (The notation (X,Y) here represents the outcome in which player x plays X and player y plays Y.) The edges of the quadrilateral are the points attainable when one player mixes and the other plays a pure strategy. The southwest edge, for example, connecting the payoffs for (C, D) and (D, D), comprises the payoffs when y plays D and x mixes between C and D. Similarly, the edge between (C, C) and (D, C) comprises the payoffs when y plays C and x mixes. A line which joins these two lines 1/4 of the way between their tops (where x plays C)and bottoms (where x plays D) comprises all the payoffs for outcomes in which x plays the (.75,.25) mix of C and D. Thus, the crossing interior lines occupy the space where both players mix. The fact that the northeast vertex of the quadrilateral is contained within this region indicates that the game is impure. We can see that the region is bounded by the edges of the quadrilateral on the southwest and southeast. The northeast boundary is a concave curve comprised of three segments. Near the top and bottom, the curve coincides with the boundaries of the underlying quadrilateral. Near the middle, the curve bends slightly to the northeast of the quadrilateral. The region northeast of the (C, C) payoff bounded by the concave curve represents the *mixing advantage*, i.e., the payoffs that both players prefer to what they get by mutual cooperation. The portion of this region along the curve represents the *optimal* mixing advantage.

A more precise characterization of the space of feasible payoffs requires two pieces of information:

- the coordinates of the two points where the optimal mixing advantage begins to depart from the boundaries of the quadrilateral

- the equation for the middle, non-linear section of the optimal mixing curve.

The second part of the task is attempted first. We describe a procedure for obtaining an equation representing the optimal mixing advantage. Neither the procedure nor the equation it yields is simple. The reader may wish to use a computer, as the author did, to verify the steps. [2]The complexity of the procedure and the fact that a computer program was used in calculating the steps involved might weaken confidence in the results. For this reason we also describe and implement an alternative method of calculating the mixing advantage. The observation that both methods lead to the same equation lends credibility to the results.

For every q, $0 \leq q \leq 1$, let us call the line representing all the points where y plays the $(q, 1-q)$ mix of cooperation and defection a *q-line*. The first method is that of *intersecting q-lines*. For $i = x, y$ let S_i, P_i, R_i and T_i be the sucker, punishment, reward and temptation payoffs to player i. Then the line segment from (C, C) to (C, D) comprises the points:

$$(qR_x + \overline{q}S_x, qR_y + \overline{q}T_y)$$

for $0 \leq q \leq 1$. (q indicates the probability with which y cooperates and $\overline{q}$ is $1-q$.) Similarly, the line from (D, C) to (D, D) comprises the points:

$$(qT_x + \overline{q}P_x, qS_y + \overline{q}P_y)$$

For each q q, $0 \leq q \leq 1$, the q-line is the line segment connecting the coordinates described by these two expressions. The equation of the line containing this segment is given by

$$\frac{y - (qS_y + \overline{q}P_y)}{x - (qT_x + \overline{q}P_x)} = \frac{(qS_y + \overline{q}P_y) - (qR_y + \overline{q}T_y)}{(qT_x + \overline{q}P_x) - (qR_x + \overline{q}S_x)}$$

Solving for y,

$$\begin{aligned} y = {} & \tfrac{1}{P_x(q-1)+S_x+q(R_x-S_x-T_x)}(P_x(q-1)(qR_y+T_y-qT_y)+q(qR_xS_y+ \\ & S_xS_y - qS_xS_y - qR_yT_x + (q-1)T_xT_y) + (q(R_y - S_y - T_y) + T_y)x + \\ & P_y(q-1)(-S_x + q(S_x - R_x) + x)) \end{aligned} \tag{1}$$

The equation for the the nearby $(q+\Delta q)$-line can be found similarly. By solving the two equations simultaneously, we obtain the coordinates of the intersection of the q-line and the $(q+\Delta q)$-line. In particular,

$$\begin{aligned} x = {} & \tfrac{1}{((R_x-T_x)(P_y-T_y)-(P_x-S_x)(R_y-S_y))}(P_yq^2R_x^2+2P_yqR_xS_x-2P_yq^2R_xS_x+ \\ & P_yS_x^2-2P_yqS_x^2+P_yq^2S_x^2-q^2R_x^2S_y-2qR_xS_xS_y+2q^2R_xS_xS_y-S_x^2S_y+ \end{aligned}$$

[2] Computations were done with *Mathematica*, making frequent use of the ”FullSimplify” command

$$\begin{array}{l}
2qS_x^2S_y - q^2S_x^2S_y - P_yq^2R_xT_x + q^2R_xR_yT_x - P_yS_xT_x + P_yq^2S_xT_x + \\
2qR_yS_xT_x - q^2R_yS_xT_x + q^2R_xS_yT_x - q^2S_xS_yT_x - q^2R_yT_x^2 + \Delta q(S_x + \\
q(R_x - S_x - T_x))((R_x - S_x)(P_y - S_y) + T_x(R_y - T_y)) + T_x(S_x - 2qS_x + \\
q^2(-R_x + S_x + T_x))T_y + P_x^2(q-1)(\Delta q + q - 1)(T_y - R_y) + P_x(-q^2R_xR_y + \\
P_y(q-1)(\Delta q + q - 1)(R_x - S_x) + R_yS_x - 2qR_yS_x + q^2R_yS_x + 2qR_xS_y - \\
q^2R_xS_y + S_xS_y - 2qS_xS_y + q^2S_xS_y - 2qR_yT_x + 2q^2R_yT_x + (q-1)(R_x + \\
qR_x + S_x - qS_x - 2qT_x)T_y + \Delta q((R_x - S_x)S_y - R_y(S_x + T_x) + (S_x + \\
T_x)T_y + q(R_yS_x + S_xS_y + 2R_yT_x - R_x(R_y + S_y - T_y) - (S_x + 2T_x)T_y))))
\end{array}$$

As Δq approaches zero, the intersection of the q- and $(q + \Delta q)$-lines approach the curve of interest to us. At the limit,

$$\begin{array}{l}
\lim_{\Delta q \to 0} x = \frac{1}{(P_x - S_x)(R_y - S_y) - (R_x - T_x)(P_y - T_y)}(-P_y(q(R_x - S_x) + S_x)^2 + \\
q^2R_x^2S_y + 2qR_xS_xS_y - 2q^2R_xS_xS_y + S_x^2S_y - 2qS_x^2S_y + q^2S_x^2S_y - \\
q^2R_xR_yT_x - 2qR_yS_xT_x + q^2R_yS_xT_x + P_y(q^2(R_x - S_x) + S_x)T_x - \\
q^2R_xS_yT_x + q^2S_xS_yT_x + q^2R_yT_x^2 + P_x^2(q-1)^2(R_y - T_y) - T_x(S_x - \\
2qS_x + q^2(S_x - Rx + T_x))T_y + P_x(P_y(q-1)^2(S_x - R_x) - S_x(R_y + \\
S_y) + (R_x + S_x)T_y + q^2(-S_xS_y - R_y(S_x + 2T_x) + R_x(R_y + S_y - T_y) + \\
S_xT_y + 2T_xT_y) + 2q(-R_xS_y + S_xS_y + R_y(S_x + T_x) - (S_x + T_x)T_y)))
\end{array}$$

This expression gives the x-coordinate of a point on the curve. It defines x as a function of q, which is one-one while $0 \leq q \leq 1$. The inverse is given by

$$\begin{array}{l}
q = \frac{1}{P_x + R_x - S_x - T_x}(P_x - S_x + \frac{1}{(R_x - S_x)(P_y - S_y) - (P_x - T_x)(R_y - T_y)} \\
\sqrt{-(-(P_x - S_x)(R_y - S_y) + (R_x - T_x)(P_y - T_y))} \\
\sqrt{(R_x - S_x)(P_y - S_y) - (P_x - T_x)(R_y - T_y)} \\
\sqrt{P_xR_x - S_xT_x + (-P_x - R_x + S_x + T_x)x})
\end{array}$$

Substituting for q in equation 1 and simplifying gives the equation for y as a function of x that was sought.

$$\begin{array}{l}
y = \frac{1}{(P_x + R_x - S_x - T_x)^2}(P_yR_x^2 + P_x^2R_y - P_yR_xS_x - R_xS_xS_y + S_x^2S_y - \\
P_yR_xT_x + 2P_yS_xT_x + 2R_yS_xT_x - S_xS_yT_x - R_xT_xT_y - S_xT_xT_y + \\
T_x^2T_y - P_x(P_yR_x + R_yS_x + S_xS_y + R_yT_x + T_xT_y + R_x(R_y - 2(S_y + \\
T_y))) + P_x(P_y + R_y - S_y - T_y)x + (R_x - S_x - T_x)(P_y + R_y - S_y - T_y)x - \\
2\sqrt{-(-(P_x - S_x)(R_y - S_y) + (R_x - T_x)(P_y - T_y))} \\
\sqrt{((R_x - S_x)(P_y - S_y) - (P_x - T_x)(R_y - T_y))(P_xR_x - S_xT_x + (-P_x - R_x + S_x + T_x)x)})
\end{array} \tag{2}$$

Let us say $y = m(x)$ for short. When we set the payoffs to $S_x = S_y = 0, P_x = P_y = 1, R_x = R_y = 3, T_x = T_y = 7$, (as was done in the illustrative graph above), the equation for the curve is given by:

$$y = x + 16\frac{1}{3} - 4\frac{2}{3}\sqrt{3 + 3x}$$

The second method for finding our equation is to compute the *highest q-line* at a given x. Consider the equation for an arbitrary q-line given in 1 above. To find the value of q at which y is maximum for a given x, we set $\frac{\partial y}{\partial q} = 0$:

$$\frac{1}{(P_x(q-1)+S_x+q(R_x-S_x-T_x))^2}((P_x(q-1)+S_x+q(R_x-S_x-T_x))(-P_xR_y+2qR_xS_y+S_xS_y-2qS_xS_y-2qR_yT_x-T_xT_y+2qT_xT_y+2P_x(q(R_y-T_y)+T_y)-(-R_y+S_y+T_y)x+P_y(R_x-2qR_x+2(-1+q)S_x+x))-(P_x+R_x-S_x-T_x)(P_x(q-1)(qR_y+T_y-qT_y)+q(qR_xS_y+S_xS_y-qS_xS_y-qR_yT_x+(q-1)T_xT_y)+(q(R_y-S_y-T_y)+T_y)x+P_y(q-1)(-S_x+q(-R_x+S_x)+x)))=0$$

Solving for q,

$$q=\frac{1}{P_x+R_x-S_x-T_x}(P_x-S_x+\frac{1}{(R_x-S_x)(P_y-S_y)-(P_x-T_x)(R_y-T_y)}\sqrt{-(-(P_x-S_x)(R_y-S_y)+(R_x-T_x)(P_y-T_y))}\sqrt{(R_x-S_x)(P_y-S_y)-(P_x-T_x)(R_y-T_y)}\sqrt{P_xR_x-S_xT_x+(-P_x-R_x+S_x+T_x)x})$$

This is the same expression for q that was obtained by the method of intersecting q-lines above. Substituting for q in equation 1, then, will produce the same equation for y as a function of x as before.

This equation coincides with the northeast frontier of the space of feasible solutions only in its middle portion. If the nearby q-lines that figure in the first method intersect only after they cross the $((C,C),(C,D))$ boundary their intersection will not be among the feasible solutions. Similarly, if the highest q-line at a given x is below the $((C,C),(C,D))$ boundary at x, that line will not extend the solution space beyond the original quadrilateral. To calculate the points at which the the curve departs from the boundaries of the quadrilateral, we need only find the points of intersection of that curve with left and right segments of the northeast boundary of the quadrilateral. The left segment segment is the line segment between (C,C) and (D,C). The line containing this segment is given by

$$\frac{y-R_y}{x-R_x}=\frac{R_y-T_y}{R_x-S_x}$$

Solving for y,

$$y=\frac{R_yS_x-R_xT_y-R_yx+T_yx}{S_x-R_x}.$$

In the case of our example,

$$y=7-\frac{4}{3}x.$$

By solving this equation and m(x) simultaneously, we can obtain the coordinates of the point at which left segment of the northeast frontier joins the middle segment. The x coordinate of this point turns out to be:

$$x_{left}=\frac{1}{P_yR_x-P_xR_y-P_yS_x-R_xS_y+S_xS_y+R_yT_x+P_xT_y-T_xT_y}(P_yR_x^2-P_xR_xR_y-P_yR_xS_x+R_xR_yS_x-R_yS_x^2-R_xS_xS_y+S_x^2S_y+R_yS_xT_x+P_xR_xT_y-R_x^2T_y+R_xS_xT_y-S_xT_xT_y),$$

which, in our example, becomes

$$x_{left}=2.$$

Similarly, the line containing the right segment has the equation

$$y = \frac{R_x S_y - R_y T_x + R_y x - S_y x}{R_x - T_x},$$

which, in our example, becomes

$$y = 5\frac{1}{4} - \frac{3}{4}x,$$

and the x-coordinate of the point at which the middle segment of the northeast frontier joins the right segment is

$$x_{right} = \frac{1}{(P_x - S_x)(R_y - S_y) - (R_x - T_x)(P_y - T_y)}(P_x R_x (R_y - S_y) + R_x^2 S_y - R_x R_y T_x - R_y S_x T_x - R_x S_y T_x + S_x S_y T_x + R_y T_x^2 + P_y R_x (T_x - R_x) + (R_x - T_x) T_x T_y),$$

which, in our example, becomes

$$x_{right} = 4\frac{1}{3}.$$

Thus in our example the northeast frontier of the feasible outcomes is described by

$$y = \begin{cases} 7 - \frac{4}{3}x & \text{if} \quad 0 \leq x < 2; \\ x + 16\frac{1}{3} - 4\frac{2}{3}\sqrt{3+3x} & \text{if} \quad 2 \leq x \leq 4\frac{1}{3}; \\ 5\frac{1}{4} - \frac{3}{4}x & \text{if} \quad 4\frac{1}{3} < x \leq 7. \end{cases}$$

Figure 2 below, shows a plot of this equation together with a plot of the other two boundaries. This yields a picture of the space of all feasible outcomes that can be compared with the original picture of figure 2 above.

3 Notable Outcomes

In the symmetric case, where $S_x = S_y, P_x = P_y, R_x = R_y$ and $T_x = T_y$, the payoffs to x and y are the same when they each employ the same probability of cooperation. In particular the payoff for cooperating with probability p is given by

$$\pi(i) = p^2 R_i + p\overline{p} S_i + p\overline{p} T_i + \overline{p}^2 P_i \tag{3}$$

The probability of cooperation that maximizes the payoffs of both players, is obtained by setting $\frac{d(\pi)}{dp} = 0$, i.e.,

$$2(p-1)P_x + S_x + 2p(T_x - S_x - T_x) + T_x = 0.$$

Solving for p,

$$p = \frac{2P_x - S_x - T_x}{2(P_x + R_x - S_x - T_x)}$$

In the case of our original example, the maximum payoff is reached when players cooperate with probability $\frac{5}{6}$. It seems reasonable to suppose that in this

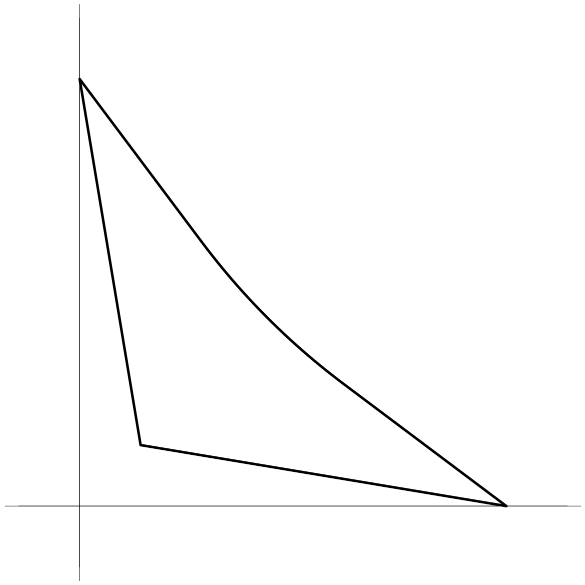

Figure 3: Feasible Solutions by Equation

case the players would agree to have that level of cooperation enforced. (Since this is a prisoner's dilemma, of course, they still have incentive to defect more frequently in the absence of enforcement.)

In asymmetric cases, it may happen that the probability of cooperation that maximizes the payoff to x may differ from the probability of cooperation that maximizes the payoff to y. In this case, it seems appropriate to relax the requirement that both players adopt the same level of cooperation. A solution that is of particular interest is one in which the mixing advantage is shared equally. This point lies on the intersection of the line with slope one through the point (R_x, R_y) and the optimal mixing advantage curve described by equation 2. The equation for the slope-one line is given by:

$$\pi(y) = \pi(x) + R_y - R_x,$$

Solving this equations and 2 simultaneously produces two ordered pairs (x_1, y_1) and (x_2, y_2), where

$$x_1 = \frac{1}{(P_x-P_y+R_x-R_y-S_x+S_y-T_x+T_y)^2}(P_x^2R_x + P_y^2R_x + R_x^3 - 2R_x^2R_y + R_xR_y^2 - 2R_x^2S_x + 3R_xR_yS_x - R_y^2S_x + R_xS_x^2 - R_yS_x^2 + R_x^2S_y - R_xR_yS_y - 2R_xS_xS_y + S_x^2S_y + S_xS_y^2 - 2R_x^2T_x + 3R_xR_yT_x - R_y^2T_x + 2R_xS_xT_x - R_xS_yT_x + R_yS_yT_x - S_xS_yT_x + R_xT_x^2 - R_yT_x^2 + (R_x^2 + R_yS_x - S_xS_y - (S_x + S_y)T_x + T_x^2 - R_x(R_y + S_x - 2S_y + 2T_x))T_y + T_xT_y^2 - 2\sqrt{(R_x - R_y + S_y - T_x)(-(P_x - S_x)(R_y - S_y) + (R_x - T_x)(P_y - T_y))} \tag{4}$$

$$\sqrt{((R_x - S_x)(P_y - S_y) - (P_x - T_x)(R_y - T_y))(R_x - R_y - S_x + T_y)}+ P_y(-S_xS_y+2S_xT_x+R_y(S_x+T_x)-T_xT_y-R_x(S_y+T_y))+P_x(-2P_yR_x+ 2R_x^2 + 2R_y^2 + R_yS_x - 2R_yS_y - S_xS_y + R_yT_x - (2R_y - 2S_y + T_x)T_y + R_x(-4R_y - 2S_x + 3S_y - 2T_x + 3T_y)))$$

$$y_1 = \frac{1}{(P_x-P_y+R_x-R_y-S_x+S_y-T_x+T_y)^2}(P_x^2R_y + P_y^2R_y + R_y^3 - P_xR_yS_x + R_y^2S_x - 2R_y^2S_y - P_xS_xS_y - 2R_yS_xS_y + S_x^2S_y + R_yS_y^2 + S_xS_y^2 - P_xR_yT_x + R_y^2T_x + 2R_yS_xT_x - R_yS_yT_x - S_xS_yT_x + R_x^2(R_y - S_y - T_y) - (2R_y^2 + (-2P_x+S_x)S_y + (P_x+S_x+S_y)T_x - T_x^2 + R_y(S_x-2S_y+2T_x))T_y + (R_y + T_x)T_y^2 - 2\sqrt{(R_x - R_y + S_y - T_x)(-(P_x - S_x)(R_y - S_y) + (R_x - T_x)(P_y - T_y))} \sqrt{((R_x - S_x)(P_y - S_y) - (P_x - T_x)(R_y - T_y))(R_x - R_y - S_x + T_y)}+ R_x(-2R_y^2 + S_y(P_x - S_y + T_x) - R_y(S_x - 3S_y + T_x - 3T_y) + (P_x + S_x)T_y - T_y^2) + P_y(2R_x^2 - 2P_xR_y + 2R_y^2 + 3R_yS_x - 2R_yS_y - S_xS_y + 3R_yT_x + 2S_xT_x - (2R_y + T_x)T_y + R_x(-4R_y - 2S_x + S_y - 2T_x + T_y)))$$

$$x_2 = \frac{1}{(P_x-P_y+R_x-R_y-S_x+S_y-T_x+T_y)^2}(P_x^2R_x + P_y^2R_x + R_x^3 - 2R_x^2R_y + R_xR_y^2 - 2R_x^2S_x + 3R_xR_yS_x - R_y^2S_x + R_xS_x^2 - R_yS_x^2 + R_x^2S_y - R_xR_yS_y - 2R_xS_xS_y + S_x^2S_y + S_xS_y^2 - 2R_x^2T_x + 3R_xR_yT_x - R_y^2T_x + 2R_xS_xT_x - R_xS_yT_x + R_yS_yT_x - S_xS_yT_x + R_xT_x^2 - R_yT_x^2 + (R_x^2 + R_yS_x - S_xS_y - (S_x + S_y)T_x + T_x^2 - R_x(R_y + S_x - 2S_y + 2T_x))T_y + T_xT_y^2 + 2\sqrt{(R_x - R_y + S_y - T_x)(-(P_x - S_x)(R_y - S_y) + (R_x - T_x)(P_y - T_y))} \sqrt{((R_x - S_x)(P_y - S_y) - (P_x - T_x)(R_y - T_y))(R_x - R_y - S_x + T_y)}+ P_y(-S_xS_y+2S_xT_x+R_y(S_x+T_x)-T_xT_y-R_x(S_y+T_y))+P_x(-2P_yR_x+ 2R_x^2 + 2R_y^2 + R_yS_x - 2R_yS_y - S_xS_y + R_yT_x - (2R_y - 2S_y + T_x)T_y + R_x(-4R_y - 2S_x + 3S_y - 2T_x + 3T_y)))$$

$$y_2 = \frac{1}{(P_x-P_y+R_x-R_y-S_x+S_y-T_x+T_y)^2}(P_x^2R_y + P_y^2R_y + R_y^3 - P_xR_yS_x + R_y^2S_x - 2R_y^2S_y - P_xS_xS_y - 2R_yS_xS_y + S_x^2S_y + R_yS_y^2 + S_xS_y^2 - P_xR_yT_x + R_y^2T_x + 2R_yS_xT_x - R_yS_yT_x - S_xS_yT_x + R_x^2(R_y - S_y - T_y) - (2R_y^2 + (-2P_x+S_x)S_y + (P_x+S_x+S_y)T_x - T_x^2 + R_y(S_x-2S_y+2T_x))T_y + (R_y + T_x)T_y^2 + 2\sqrt{(R_x - R_y + S_y - T_x)(-(P_x - S_x)(R_y - S_y) + (R_x - T_x)(P_y - T_y))} \sqrt{((R_x - S_x)(P_y - S_y) - (P_x - T_x)(R_y - T_y))(R_x - R_y - S_x + T_y)}+ R_x(-2R_y^2 + S_y(P_x - S_y + T_x) - R_y(S_x - 3S_y + T_x - 3T_y) + (P_x + S_x)T_y - T_y^2) + P_y(2R_x^2 - 2P_xR_y + 2R_y^2 + 3R_yS_x - 2R_yS_y - S_xS_y + 3R_yT_x + 2S_xT_x - (2R_y + T_x)T_y + R_x(-4R_y - 2S_x + S_y - 2T_x + T_y)))$$

To find the probabilities of cooperation that x and y must employ to achieve these payoffs, observe that the payoffs to x and y when x cooperates with probability p and y cooperates with probability q are given by

$$\pi(x) = pqR_x + p\overline{q}S_x + \overline{p}qT_x + \overline{pq}P_x \text{ and}$$

$$\pi(y) = pqR_y + p\overline{q}T_y + \overline{p}qS_y + \overline{pq}P_y.$$

Solving $x_1 = \pi(x)$ and $y_1 = \pi(y)$ simultaneously, we obtain

$$\begin{array}{l}
p = \frac{1}{(-(P_x-S_x)(R_y-S_y)+(R_x-T_x)(P_y-T_y))(P_x-P_y+R_x-R_y-S_x+S_y-T_x+T_y)}((P_x- \\
P_y + S_y - T_x)(-(P_x - S_x)(R_y - S_y) + (R_x - T_x)(P_y - T_y)) + \\
\sqrt{(R_x - R_y + S_y - T_x)(-(P_x - S_x)(R_y - S_y) + (R_x - T_x)(P_y - T_y))} \\
\sqrt{((R_x - S_x)(P_y - S_y) - (P_x - T_x)(R_y - T_y))(R_x - R_y - S_x + T_y)}),
\end{array}$$

$$\begin{array}{l}
q = \frac{1}{P_x-P_y+R_x-R_y-S_x+S_y-T_x+T_y}(P_x-P_y-S_x+T_y-\frac{1}{-(R_x-S_x)(P_y-S_y)+(P_x-T_x)(R_y-T_y)} \\
\sqrt{(R_x - R_y + S_y - T_x)(-(P_x - S_x)(R_y - S_y) + (R_x - T_x)(P_y - T_y))} \\
\sqrt{((R_x - S_x)(P_y - S_y) - (P_x - T_x)(R_y - T_y))(R_x - R_y - S_x + T_y)}
\end{array}$$

For example, if the payoffs for x are 0,1,3 and 7, as in the original example, and those for y are .5,2,4 and 9, then this solution calls for x and y to cooperate with probabilities 80.5% and 81.5%, respectively, giving them payoffs of 3.12 and 4.12. Each gets the same .12 advantage from mixing.

4 Impurity in Other Games

In an impure prisoner's dilemma there is a mixed strategy pair that affords both players a higher expected payoff than the "natural" or "desirable" outcome of (C, C). A similar phenomenon can occur with other games. Consider, for example, the three games with matrices below. (The moves have been labelled C and D for ease of comparison with the prisoner's dilemma.)

	C	D
C	3,3	2,7
D	7,2	0,0

	C	D
C	0,0	2,7
D	7,2	3,3

	C	D
C	0,0	2,7
D	7,2	1,1

Figure 4: Three Impure Games

Note that the first two games are distinct in the sense that neither can be obtained from the other by merely relabelling players or moves. Nevertheless the *diagrams* of the two games (in the sense that the quadrilateral of figure 1 is the diagram of the prisoner's dilemma of our earlier example) are the same, and they are similar to the diagrams of the third game and the impure prisoner's dilemma. The first game is a version of Chicken, in which C and D are *Dove* and *Hawk*, respectively. The third game can be regarded as a version of Luce and Raiffa's Battle of the Sexes where, for each player, to play C is to attend the event favored by the other and to play D is to go to the event favored by itself. The version envisioned here substitutes the *liberal* assumption that both parties prefer the outcome where each attends its preferred event alone to the outcome where each attends at its less preferred event alone for the *romantic* assumption that they are indifferent between these outcomes. The second game can be regarded as an even less romantic version of Battle of the Sexes. Each player here prefers attending the preferred event alone to attending the less preferred event together.)

In the chicken example, there are two pure nash equilibria, (C, D) and (D, C), and a mixed equilibrium in which both players mix $\frac{1}{3}C$ with $\frac{2}{3}D$. For some applications (for example in biological models which take the strategy to be a part of the genetic makeup of members of the species), outcomes in which the players play different strategies are deemed unattainable, and so the mixed strategy is the game's sole "solution." Since the payoff to the mixed strategy to each player is only $2\frac{1}{3}$ and each player would get 3 by mutual cooperation this gives the game the feel of a prisoner's dilemma. It makes sense in this case to extend our terminology. A chicken game with mixed strategies that provide a greater expected payoff to both players than mutual cooperation can be said to be impure. If we label the payoffs to pure strategy pairs as in the prisoner's dilemma then, when $R_x > S_x$ and $R_y > S_y$, condition I above provides necessary and sufficient conditions for a chicken game to be impure. (Any chicken game in which $R_x < S_x$ and $R_y < S_y$ is also impure.) For example, in the chicken example above, both players can get an expectation of $3\frac{1}{9}$ if they each mix $\frac{2}{3}C$ with $\frac{1}{3}D$, compared with $2\frac{1}{3}$ for their mixed equilibrium and 3 for mutual cooperation.

Similarly, in the last game, (C, D) and (D, C) are pure strategy equilibria (representing the outcomes where the couple goes to the same event) and there is one mixed strategy equilibrium where both players mix at $\frac{1}{4}C$ and $\frac{3}{4}D$. The mixed strategy equilibrium gives each player a payoff of $1\frac{3}{4}$. In contrast to the chicken game, they both prefer this outcome to mutual cooperation and mutual defection. Again, however, there are other independent mixes that allow them both to do even better. If both mix at $\frac{3}{4}C, \frac{1}{4}D$, for example, they each achieve 2 units. If we regard the mixed equilibrium as a natural or desirable outcome (for example, because considerations of fairness count against the other two equilibria) then it seems reasonable to extend the notion of impurity) to cover this case of the liberal Battle of the Sexes. The case for the second example is even stronger. Here, mutual defection, yielding a payoff of 3 units for each player, is the sole equilibrium, yet by mixing at $\frac{1}{6}C, \frac{5}{6}D$, both players get an expected payoff of $3\frac{1}{3}$. So this is another example of an impure game.

In Prisoner's Dilemmas and Chicken games, the point of mutual cooperation lies northeast of the point of mutual defection. In the Liberal Battle of the Sexes games the point of mutual defection lies northeast of the point of mutual cooperation. Thus the impurity condition, that (R_x, R_y) lies to the southwest of the line between (S_x, T_y) and (S_y, T_x), should be replaced by a condition that (P_x, P_y) lies to the southwest of that line. This implies that condition I should be replaced by

$$I') \quad (T_x - P_x)(T_y - P_y) < (P_x - S_x)(P_y - S_y)$$

(This condition presumes that $P_x > S_x$ and $P_y > S_y$. If the genus of games includes cases in which, say, $P_x < S_x$ and $P_y < S_y$, then all such cases would be impure.)

The preceding discussion might suggest that any game whose diagram is concave on the northeast side be regarded as impure. The following two examples, however, should give us pause.

	C	D
C	5,1	2,7
D	7,2	1,5

	C	D
C	8,0	2,7
D	7,2	0,8

Figure 5: Two Pure Games?

In both cases the sole nash equilibrium is (C, D), and in both cases no outcome (mixed or pure) affords both players a greater expected payoff than this outcome. It is true that in the second game there are pairs of mixed strategies (e.g., $\frac{1}{6}C, \frac{5}{6}D$ for x and $\frac{5}{6}C, \frac{1}{6}D$ for y) that afford both players a higher expected payoff than (D, D) and others (e.g., $\frac{5}{6}C, \frac{1}{6}D$ for x and $\frac{1}{6}C, \frac{5}{6}D$ for y) that afford both players a higher expected payoff than (C, C). It is hard to imagine, however, an interpretation in which either (C, C) or (D, D) would be regarded as the natural or desirable outcome for this game. And in the second example, because R_y and P_x are the largest payoffs in the matrix, no outcome, mixed or pure affords both players a higher expectation than (C, C) or (D, D).

The notion of impurity given here, resting as it does on the notions of *natural* or *desirable* outcomes, is an *informal* notion and one which might, in some cases, depend on the intended application of the game as well as the game itself. Alternatively, impurity could have been identified with some *formal* property that generalizes on the impure prisoner dilemmas. The most likely candidate is the property that, for every pure outcome of the form (X,X), there is some pair of independent mixed strategies that provides both players a greater expected payoff. In that case we would would say that the second of the two examples in figure 4, in addition to the three examples of figure 3 are impure, and only the last example of figure 4 is not. I prefer to keep the label to describe what seems to me to be the more interesting examples.

References

[1] Gauthier D. Morality and Advantage. *Philosophical Review* **LXXVI (3)** pp460-475, 1967.

[2] Kuhn, S. T. Agreement Keeping and Indirect Moral Theory. *Journal of Philosophy* **XCIII(3)** pp105-128, 1996.

[3] Kuhn, S.T. and S. Moresi. Pure and Utilitarian Prisoner's Dilemmas. *Economics and Philosophy* **11,** pp123-133, 1995.

[4] Moresi, S and S.T. Kuhn. Pure and Utilitarian Prisoner's Dilemmas. *Working Papers 94-03.* Georgetown University Department of Economics, Georgetown University, Washington DC 20057, 1994.