On the Impossibility of Prisoner's Dilemmas in Adversarial Preference Games

with
Implications for Conflict Inducing Growth

****
May 25, 2003
Preliminary First Draft
(comments solicited)
****

JEL C7, D6, H4

By

Andrew W. Horowitz

Department of Economics
Sam M. Walton College of Business
University of Arkansas, Fayetteville, AR 72701

Abstract: Why don't agents cooperate when they both stand to gain? This question ranks among the most fundamental in the social sciences. Explanations abound. Among the most compelling are various configurations of the prisoner's dilemma (PD), or public goods problem. Payoffs in PD's are specified in one of two ways: as primitive cardinal payoffs *or* as ordinal final utility. However, as final utility is objectively unobservable, only the primitive payoff games are ever observed. This paper explores mappings from *primitive* payoff *to utility* payoff games and demonstrates that though an observable game is a *PD* there are broad classes of utility functions for which there exists no associated *utility* PD. In particular we show that even small amounts of either altruism or jealousy may disrupt the mapping from *primitive* payoff to *utility* PD. We then examine some implications of these results – including the possibility of conflict inducing growth.

Horowitz email. Horowitz@walton.uark.edu; Phone: 479-575-6228. I thank Arilton Teixeira, Aloisio Pessoa de Araújo, Amy Farmer, and Joel Carron for helpful comments.

I. Introduction

Prisoner dilemmas (PDs) have been employed across the social and business sciences, philosophy, and biology as prime examples of the tension between individual and collective rationality.[1] They constitute powerful illustrations of the gains foregone when strategic structure precludes cooperation as an equilibrium strategy. The payoffs in PD's have two forms. First, they may be cardinal observable payoffs (e.g., years in prison, nuclear warheads, or advertising budgets). We refer to such games as *Primitive Prisoner's Dilemmas* (PPDs). Alternatively, payoffs may be specified as final utility, which is inherently unobservable. We refer to these games as *Utility Prisoner's Dilemmas* (UPDs). In either case there is an implicit mapping between observable payoff and final utility that has received scant attention in the literature. Though this neglect may be innocuous for some mappings from payoff to utility we demonstrate in this paper that it may substantive for broad classes of utility mapping. Specifically, we identify broad classes of utility mappings for which it is impossible to map a PPD into a UPD. These include the case where players' preferences reflect either an adversarial or altruistic inclination towards the other player.

Much of this manuscript focuses on adversarial preference mappings from primitive payoff to utility. With adversarial preferences a player's utility is strictly decreasing in rivals' primitive payoff, all else equal.[2] Adversarial relationships, in the

[1] A nice survey of economic applications of the PD can be found in Rapoport 1987.

[2] We wish to emphasize the distinction between adversarial *payoff structure* – such as zero sum games -- and the adversarial *preferences* that are our focus

sense of competition, arise in virtually all economic environments. However, it is typically assumed that own-utility is independent of other players' payoffs, all else equal. In this sense players are *strategic adversaries* in the typical competitive environment in that their incentive to adopt a particular strategy is governed by own payoff maximization rather than explicit consideration of rivals' payoff.[3] When rival payoff adversely affects own-utility, we call rivals *absolute adversaries.* Such preferences may correspond to conventional notions envy or malice. These terms -- "envy" and "malice" -- have precise economic meanings (see for example Hammond 1987; Chaudhuri 1985; and Brennan 1973), and though this literature addresses issues tangentially related to this paper, none of the literature addresses the implications of such utility mappings on the existence of the PD.

Another motivation for the preference structure we explore is a game where joint strategies map into two-good payoffs. Adversarial preferences are technically equivalent to the case where one of the primitive payoffs is a "good" and other is a "bad" for one player, while the second player has reverse preferences towards the payoffs. For example, we can imagine roommates who have contradictory preferences towards classical and rock music. For one of the occupants classical is a good and rock is a bad, while the reverse holds for the other roommate. Joint strategies yield quantities of both goods, and it is easy to construct PD structure (i.e., Pareto Inferior equilibrium) in this environment. Though the narrative of this manuscript focuses on adversarial preferences, some may prefer the good-bad interpretation.

[3] In zero sum games these objectives would be equivalent. But as noted, our analysis does not concern zero-sum games.

Our analysis of adversarial mappings generates two results with a number of important implications. The first result is that for a broad class of such adversarial mappings it is *impossible* for a PD in observable payoffs to map into a *utility PD*. A second result concerns the class of adversarial utility functions for which *it is possible* to have a utility PD associated with a payoff PD. In this case we demonstrate a non-congruence of the primitive-payoff and utility games. An implication of this second result is that when the payoffs of all players grow proportionally a *utility-PD* that is associated with a *payoff-PD* must eventually cease to exist. Thus, any cooperative mechanism designed to overcome the Pareto inferior PD equilibrium will be undermined by growth itself. This raises the possibility of conflict inducing growth. We also briefly consider altruistic preferences in the *UG*. Though this section is preliminary, we show that the PPD need not map to a UPD in the transformed game.

II. Analysis

II.1 Notation and Definitions

Consider a two-player game and call the players A and B and their cardinal (observable) payoffs α and β respectively. Each player has two strategies. Denote the players strategy sets and strategy choice as respectively: $S^p = \{1, 2\}$ and s^p for $p = A, B$. So the joint strategy space has four elements and denote the associated observable primitive payoff vector as $\pi_{ij} = [\alpha_{ij}, \beta_{ij}]$ where $i = s^A$ and $j = s^B$. When the clarity constraint permits we suppress the subscripts on α and β.

Each player has unobservable preferences over the primitive payoff space that are complete, transitive, and reflexive. In a slight (but innocuous) abuse of notation that yields considerable notational economy we denote the unobservable utility functions as:

$A(\alpha, \beta)$, $B(\alpha, \beta)$. We will initially assume that these utility functions are differentiable in both arguments, but will subsequently demonstrate that all our results hold if we relax differentiability and retain continuity.

Definitions

The following definitions will facilitate derivation and exposition of our results.

(i). Primitive Game (PG)

We define a *Primitive Game* as $\Gamma = [P, S, \Pi]$ with P the player set, S the strategy set, and Π the payoff set. Π is a set of primitive observable objects, rather than ultimate utility. In the primitive game, observable own-payoffs are a perfect proxy for final utility with more own-primitive-payoff always preferred to less regardless of rival's payoff.

(ii). Primitive Prisoners Dilemma (PPD)

A *PPD* requires the existence of primitive joint payoff that vector dominates another joint payoff. Without loss of generality let $s^i = 1$ for $(i = A, B)$ be the strategies that map to the Pareto Inferior payoff vector and $s^i = 2$ for $(i = A, B)$ the strategies that map to the Pareto Superior payoff vector. *A* primitive game, Γ, is a PPD if the payoffs satisfy the following conditions: *(i).* $\pi_{22} >> \pi_{11}$. *(ii).* $\alpha_{12} > \alpha_{22}$; *(iii).* $\alpha_{21} < \alpha_{11}$; *(iv).* $\beta_{12} < \beta_{11}$; . *(v).* $\beta_{21} > \beta_{22}$. Given relationships *(i) - (iv)* the unique Nash Equilibrium is: $s^i = 1$ for $(i = A, B)$, which yields Pareto Inferior primitive payoffs of π_{11}. Call the set of all PPDs, Γ^{PD} with $\Gamma_i^{PD} \in \Gamma^{PD}$.

(iii). Primitive Superior Set (PSS)

Consider an arbitrary $\pi_{11.}$ Define the *Primitive Superior Set (PSS)* as: $\Pi_{22}(\pi_{11}) = \{\pi >> \pi_{11}\}$. In the payoff space the set Π_{22} is the quadrant to the northeast of π_{11}. That is, the set of primitive payoff vectors that are Pareto Superior to π_{11}.

(iv). Primitive Dominant Sets (PDS^i)

Given an arbitrary π_{11} and $\pi_{22} \in \Pi_{22}(\pi_{11}).$ Define the *Primitive Dominant Set* (PDS^i) of player $\{i = A, B\}$ as respectively: $\Pi_{12}(\pi_{11}) = \{\pi\text{: } \alpha > \alpha_{22}, \beta < \beta_{11}\}$ and $\Pi_{21}(\pi_{11}) = \{\pi\text{: } \alpha < \alpha_{11}, \beta > \beta_{22}\}$. The PDS of player i, (denoted Π_{ij}) is thus the set of primitive payoffs that dominate π_{22} for player i and are dominated by π_{11} for player j.

Figure 1a below illustrates the PSS and PDS's in the primitive payoff space. An arbitrary π_{11} together with any triplet $\{\pi_{22} \in \Pi_{22}(\pi_{11}), \pi_{12} \in \Pi_{12}, \pi_{21} \in \Pi_{21}\}$ yield a PPD (Primitive Prisoners Dilemma). Whenever the triplet falls in the shaded areas indicated in Figure 1a, a PPD occurs.

(v). Utility Game (UG)

Let $U_{ij}= [A(\pi_{ij}), B(\pi_{ij})]$. So U_{ij} is a vector of final utilities from primitive payoffs when player A plays strategy i and B plays strategy j (where i may equal j). At this point we place no restrictions on U. Therefore, the *UG* admits classes of preference mappings not allowed in the primitive game. Specifically, the functions $A(\pi_{ij})$ and $B(\pi_{ij})$ may map non-monotonically from own-primitive-payoff to final utility due either adversarial or altruistic preferences. Every *Primitive Game* maps to an associated *Utility Game* (*UG*) for a given U. Thus, given a primitive game, Γ, and a preference map U, we define the associated *UG* as $V(\Gamma) = [P, S, U(\Pi)]$. Naturally, if U does not have the direct correspondence property of the primitive game (implicit) preferences, V will be a better predictor of players' strategic behavior than Γ.

(vi). Utility Prisoner Dilemma (UPD)

A utility game, $V(\Gamma)$ is a *UPD* iff: *(i).* $A(\pi_{22}) > A(\pi_{11})$; *(ii).* $B(\pi_{22}) > B(\pi_{11})$; *(iii).* $A(\pi_{12}) > A(\pi_{11})$; *(iv).* $A(\pi_{21}) < A(\pi_{11})$; *(v).* $B(\pi_{12}) < B(\pi_{11})$; *(vi).* $B(\pi_{21}) > B(\pi_{22})$. That is, the unique Nash Equilibrium is Pareto inferior in utility payoffs. Let V^{PD} denote the set of all UPDs, with $V_i^{PD} \in V^{PD}$.

(vii). Utility Pareto Superior Set (USS)

Consider an arbitrary π_{11}. Define the *USS* as: $U_{22}(\pi_{11}) = \{\pi\text{: } A(\pi) > A(\pi_{11}),\ B(\pi) > B(\pi_{11})\}$. . Note that *USS* is a set the primitive joint payoff space and also note that $U_{22}(\pi_{11})$ may be an empty set.

(viii). Utility Dominant Sets (UDS^i)

Given an arbitrary π_{11} and $\pi_{22} \in \Pi_{22}(\pi_{11})$. Define the *Utility Dominant Set (UDS^i)* of player $\{i = A, B\}$ as respectively: $U_{12}(\pi_{11}) = \{\pi\text{: } A(\pi) > A(\pi_{22}),\ B(\pi) < B(\beta_{11})\}$ and $U_{21}(\pi_{11}) = \{\pi\text{: } A(\pi) < A(\pi_{11}),\ B(\pi) < B(\pi_{22})\}$. UDS^i is thus the set primitive payoffs that *utility* dominate π_{22} for player *i* and are *utility* dominated by π_{11} for player *j*.

Corresponding Primitive and Utility Prisoners Dilemmas

Suppose a Primitive Game is a PD. What can we say about the existence and nature of the associated UG? More specifically, if we admit the possibility of some adversarial or altruistic component to preferences, under what classes of utility functions will the associated UG also be a PD?

To answer this question we begin as a point of reference with the standard assumption in PDs of no adversarial or altruistic preferences. That is, each players' utility is increasing in own primitive payoff and independent of rival's primitive payoff (for a given strategy). Letting subscripts denote partials in the usual manner:

$A_\alpha > 0,\ A_\beta = 0,\ B_\beta > 0,\ B_\alpha = 0.$ This generates the linear indifference curves in the joint payoff space, as illustrated in Figure 1b below. Note that is this standard case, for any arbitrary π_{11} and an $\pi_{22} \in \Pi_{22}(\pi_{11})$ the PSS = USS and the PDS^{i} = UDS^{i} for $i = (A, B\}$. Thus every primitive prisoners' dilemma is also a utility prisoners dilemma. Consequently, in the absence of adversarial or altruistic preferences there is no loss in predictive power when the UG is ignored and all attention is directed to the primitive game.

.

[Figure 1a and 1b go here]

II.2 Adversarial Preferences

Now suppose that, in contrast to the usual assumption, *B*'s payoff affects *A*'s utility *adversely* and *A*'s payoff affects *B*'s utility *adversely,* all else equal.[4] Then the players preferences exhibit the following properties with respect to the others player's primitive payoffs: $A_\alpha > 0,\ A_\beta < 0$ and $B_\beta > 0,\ B_\alpha < 0.$ Consequently the indifference curves are upward sloping in the joint payoff space. For expositional simplicity we will focus on the case where these indifference curves have finite slope. We begin with some general results on the relationship between the sets defined above. We then consider three classes of adversarial preferences and derive more specific existence results.

[4] As noted, this is technically equivalent to α being a good for *A* and a bad for *B*, and β being a good for *B* and a bad for *A*.

Some General Results

Proposition 1. With Adversarial Preferences there exist Primitive Prisoner's Dilemmas that have no corresponding Utility Prisoner's Dilemma. Formally,

$\exists\, \Gamma_i^{PD} = [P, S, \Pi_i] \in \Gamma^{PD}$ such that $V(\Gamma_i^{PD}) = [P, S, U(\Pi_i)] \notin V^{PD}$.

Proof. Let $\Pi_i = [\,\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}]$ be a PPD quadruple. So $\pi_{22} \in \Pi_{22}(\pi_{11}) >> \pi_{11}$. With adversarial preferences indifference curves have finite upward slope in the joint payoff space. Therefore we can have $\pi_{22} \in \Pi_{22}(\pi_{11})$ and $A(\pi_{22}) = A(\pi_{11})$ so that $\pi_{22} \notin$ USS and $V(\Gamma_i^{PD}) = [P, S, U(\Pi_i)] \notin V^{PD}$ $\therefore$

Proposition 1 demonstrates that the standard implicit assumption that a game with PD structure in observable payoffs maps to a PD in utility is not necessarily true when one player has any level of adversarial preferences. We will subsequently identify a broad class of adversarial preference for which *it is impossible* for a PPD to map to UPD. Note that Proposition 1 could be made stronger by replacing π_{22} with $\pi_{22}' = \pi_{22} - \varepsilon$. Then $\pi_{22}' \in \Pi_{22}(\pi_{11})$ and $A(\pi_{22}') < A(\pi_{11})$.

Proposition 2. With adversarial preferences each player's Primitive Dominant Set is a strict sub-set of their Utility Dominant Set: $\text{PDS}^{\text{i}} \subset \text{UDS}^{\text{i}}$.

Proof. Begin with $i = A$. Again let $\Pi_i = [\,\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}]$ be a PPD quadruple and define $\pi' = [\,\alpha_{22}, \beta_{11}]$. Then $A(\pi_{22}) < A(\pi') < A(\pi \in PDS^A)$ since $\beta_{11} < \beta_{22}$ and $\forall\, \pi(\alpha, \beta) \in PDS^A$ $\alpha > \alpha_{22}$ and $\beta < \beta_{11}$. So π'. is an element of UDS but not PDS and every element of PDS must an element of UDS since $A(\pi_{22}) < A(\pi \in PDS^A)$. A similar argument holds for player B $\therefore$

Proportional Adversarial Preferences

Suppose preferences over primitive outcomes have the following form: $A(\alpha, \beta) = f\left(\frac{\alpha}{\beta}\right)$, $B(\alpha, \beta) = g\left(\frac{\beta}{\alpha}\right)$, where $f' > 0, f'' \leq 0, g' > 0, g'' \leq 0$. Thus, own-utility depends only on the ratio of own to adversary's payoffs. We will discuss settings where such preferences may be reasonable descriptions of adversarial preferences in the following section. Given the monotonicity of the *f* and *g* functions, they may be inverted and the indifference curves are rays from the origin in the primitive payoff space. Figure 2 below illustrates a PPD with Proportional Adversarial Preferences superimposed.

[Figure 2 here]

Inspection of Figure 2 reveals that for any initial π_{11} the USS is empty and therefore that a PPD can never can never have an associated UPD. To see this note that beginning anywhere on the indifference curve passing through π_{11} any clockwise movement decreases *A's* utility and increases *B's*. Counterclockwise movement has the reverse effect. Therefore the USS is empty. In the following sub-section, we will provide a formal mathematical proof of a more general case that encompasses Proportional Adversarial Preferences as a special case.

Convex Single-Crossing Indifference Curves

With Proportional Adversarial Preferences a player's utility is independent of absolute payoff level. There are many instances, however, where an adversarial mapping from payoff to utility may depend on absolute primitive payoffs. For example, consider the case where the payoffs are the GDPs of two adversarial counties. Each country's fear of their adversary may have an absolute as well as a relative component: as both countries become bigger, the absolute damage they could inflict upon each other increases. Formally, there are two requirements needed to capture these types of preferences. First, the (upward sloping) indifference curves must be convex. Second, utility must decline monotonically as we move outward along a ray from the origin. That is, ever increasing own-payoffs are required for indifference as the rival's absolute payoff increases. We capture both of these requirements in a single condition that requires the MRS_{ij} must exceed the ratio of i to j, where $i, j = \{\alpha, \beta\}$ and $i \neq j$.. Noting that with α on the vertical axis the slopes of the indifference curves for players A and B are respectively $-A_\beta/A_\alpha$ and $-B_\beta/B_\alpha$, inequalities *(1)* captures these conditions for both players.

(1) $$-A_\beta/A_\alpha > \alpha/\beta, \qquad -B_\beta/B_\alpha < \alpha/\beta.$$

These inequalities also have the geometric interpretation that at any point along a ray from the origin in the payoff space (with own-payoff on the vertical axis), the slope of the indifference curve exceeds the ray. They also constrain all indifference curves to emanate from the origin. In doing so they describe preferences under which a player is indifferent between any own-payoff so long as the player's adversary receives a null payoff. This case is therefore a generalization of the proportional utility mapping in the prior sub-section since preferences depend on absolute as well as relative payoffs. However, when the rival receives a zero payoff the relative component is explosive, and payoff vectors are indistinguishable. An interpretation is that with a zero payoff the rival is completely powerless and malice, envy, or fear evaporate. Under these conditions, as embodied in inequality *(1)*, the following proposition holds.

Proposition 3. When players are absolute adversaries a PPD will never have an associated UPD.

Proof: From our definition of a *UPD* we have the following two inequalities: *(i).* $A(\pi_{22}) > A(\pi_{11})$; *(ii).* $B(\pi_{22}) > B(\pi_{11})$. For π_{22} close to π_{11} this implies: *(iii).* $A_{\alpha}(\pi_{11})d_{\alpha} + A_{\beta}(\pi_{11})d_{\beta} > 0$; *(iv).* $B_{\alpha}(\pi_{11})d_{\alpha} + B_{\beta}(\pi_{11})d_{\beta} > 0$. *Combining (iii) and (iv) yields:* $A_{\alpha}/A_{\beta} < B_{\alpha}/B_{\beta}$. *Combining inequalities (1) above yields:* $A_{\alpha}/A_{\beta} \geq B_{\alpha}/B_{\beta}$, *a contradiction* $\therefore$

Figure 3 below illustrates this proof graphically. Note that since preferences are complete, there are indifference curves that intersect any non-null equilibrium payoff (π_{11}) in the configuration illustrated. Then in the quadrant to the northeast of π_{11}, where π_{22} must lie for a PPD to exist, either one player is worse off or both are worse off (the shaded region indicates the zone where both players have lower *utility*).

[Figure 3 here]

Multiple-Crossing Convex Indifference Curves (Potentially)

As noted previously, the behavioral assumption that for a given α/β utility declines monotonically in absolute rival payoff (as embodied in inequalities *(1))* constrained the indifference curves to emanate from the origin. The interpretation was a dominance in the preference ordering of relative over absolute positions vis-à-vis the rival. Now suppose we relax this condition and merely insist on convexity (strict) of the indifference curves (from the orientation of own-payoff on the vertical axis). That is:

$$(2) \qquad A_{\alpha\beta}A_{\beta} - A_{\beta\beta}A_{\alpha} > 0, \qquad\qquad B_{\beta\alpha}B_{\alpha} - B_{\alpha\alpha}B_{\beta} > 0.$$

These conditions ensure convexity but do not constrain the indifference curves to emanate from the origin. With these conditions *A's* utility level may increase monotonically along the vertical axis and *B's* utility increases monotonically along the horizontal axis.

Having relaxed the single crossing condition a number of interesting possibilities arise. Among these is the possibility that a sufficiently large increase in the payoff to both players may result in a regime shift from the zone where Pareto improving strategy combinations exist, to one in which no Pareto improving strategies combinations exist. To see this, first consider payoff π_{11} in Figure 4 below. Since preferences are complete any arbitrary payoff (π_{11}) is an intersection of *some A* and *B* indifference curve, as illustrated. Then the shaded area in the "lens" is a zone of Pareto Superior utility payoffs. If π_{22} falls in this zone, and if π_{12} and π_{21} fall in the PDS zones identified in Figure 1a, the payoff-PD maps into a utility-PD.

[Figure 4 here]

Now suppose that *all* payoffs increase by *x*% due to some extra-strategic growth in the system. Call the new payoffs $\hat{\pi}_{11}$ and $\hat{\pi}_{22}$. Then because π_{22} vector dominates π_{11} the absolute distance between $\hat{\pi}_{11}$ and $\hat{\pi}_{22}$ will increase. For a big enough percentage growth, the new payoff configuration will have $\hat{\pi}_{22}$ fall in the shaded area in Figure 4 labeled Pareto-Inferior. Thus, given this preference structure there always exists some payoff growth at which the association of a primitive-PD with a utility-PD collapses. Any mechanism designed to support the Pareto superior outcome before the growth will now collapse. It is this principle that generates a mechanism for *conflict inducing growth*.

II.1.c. Altruism

This section presents some preliminary discussion of the correspondence between PPD and UPD when players' preferences reflect altruism towards the other player. We emphasize that the discussion of this section is preliminary and suggestive.

Strategic models are often applied to environments where altruism is likely to be present (e.g., intra-household models) – though its presence (altruism) in the strategic environment raises a number of subtle conceptual issues that are beyond the scope of this paper. Our objective here is simply to use the set theoretic characterization of PDs we have developed to briefly explore the correspondence of PPD and UPD when preferences embody some altruism towards the other player.

With altruism, both *own* and *other* players payoffs are "goods," though we would expect them to be imperfect substitutes in the absence of some form of perfect altruism. With some altruism, if all strategies yielded the same own-payoff a strategy that provided a higher payoff to the other player would be preferred, rather than the standard assumption of indifference. The utility function partials would now be

$A_\alpha > 0,\ A_\beta > 0,\ B_\beta > 0,\ B_\alpha > 0$ and indifference curves are downward sloping in the joint payoff space. Using the standard no-altruism-preferences as a point of reference (horizontal indifference curves for *A*), as player *A* becomes altruistic towards player *B* the indifference curve exhibits some curvature, becoming steeper as the degree of altruism increases. Is it possible to have a UPD with altruistic preferences? Figure 5 below illustrates that it is indeed possible to have UPD. Perhaps more importantly Figure 5 demonstrates the non-congruence of the PPD and UPD when the players have some altruism towards each other.

[Figure 5 here]

The USS in Figure 5 is the yellow shaded area beyond the A_1 and B_1 indifference curves while the PSS remains the quadrant to the northeast of π_{11}. Proposition 1 implied that with adversarial preferences USS $\subset$ PSS. With altruism it is evident from Figure 5 that the opposite is true: PSS $\subset$ USS. Now consider the relationship between PDS^i and UDS^i. Proposition 2 established that with adversarial preferences PDS^i $\subset$ UDS^i. With altruism Utility Dominant Sets become the green shaded areas in Figure 5 (indicated by U(Π_{ij}) while the PPS remains the rectangle indicated by the Π_{ij}. Therefore, with altruism Proposition 2 is reversed and UDS^i $\subset$ PDS^i. The consequence of these relationships is that with altruism, as was the case with adversarial preferences, a Primitive Prisoners Dilemma need not imply a prisoners dilemma in final utility (UPD). Indeed, Figure 5 reveals that primitive games that do not have PD structure (Pareto inferior unique Nash equilibrium) may in fact have PD structure in the utility space.

III. Summary and Conclusion

Prisoner's dilemmas provide a fundamental paradigm of the tension between individual and collective rationality. Analysis of their structure and operation has provided insight into issues ranging from the public goods problem to arms races. Yet the predictive power of the paradigm depends critically on implicit assumptions on the nature of the mapping from observable primitive payoff to unobservable final utility. When unobservable final utility depends only on own-primitive-payoff the equilibrium of a primitive-payoff-game and the associated utility-games are identical. Under this circumstance the specific properties of the unobservable utility function are immaterial for predictions of strategy choice and a primitive game with a PD equilibrium is a perfect proxy for the unobservable final utility game. However, when linkages exist between the primitive payoff of one player and the utility of another, PD equilibrium in the observable game may not correspond to equilibrium in the utility game.

This paper explores the implications of two types of linkages between the players' final utility and the other player's primitive payoff: adversarial and altruistic preferences. There are a number of motivations and interpretations for the adversarial preference structure. Jealously or envy are natural explanations for adversarial preferences. Alternatively, all the adversarial model results are fully applicable to the case where primitive payoffs are two distinct goods: one of which enters the utility function as a good and the other as a bad. With either interpretation of adversarial preferences, we demonstrate that the linkage between the observable primitive game and the utility game is disrupted. Specifically, we demonstrate a non-congruence of the “primitive-superior-set” and “utility-superior-set” and between the “primitive-dominant-set” and “utility-dominant-sets.” A reverse non-congruence arises with altruism. The effect of these non-congruencies is that PD structure in the observable primitive payoff game does not guarantee PD structure in the utility game. Moreover, utility PDs may arise in games that do not exhibit PD structure in primitive payoffs.

To appreciate the implications of a non-congruence between the observable primitive-game and the utility-game consider a standard two-person PD in observable payoffs. Both players are *apparently* better off through cooperation than competition, though the temptation of defection precludes cooperation as a non-cooperative equilibrium. Given this well-known configuration, it would seem that the players have incentive to create institutions that can support the Pareto-superior strategy payoffs. Indeed, we observe many situations where institutions supporting the Pareto Superior outcome are created and the gains from cooperation can be realized. However, we also observe many situations where cooperation does not occur, where institutions we can easily envision are not developed, and where all players appear to reap inefficiently low returns. This paper proposes a new explanation for such phenomenon. Namely, that the joint-strategy Pareto superior *utility-payoffs* do not in fact exist. On the flip side, equilibrium which appear Pareto optimal in primitive payoffs may in fact be PDs in utility payoffs.

The current manuscript is a first exploration of the mapping between Primitive and Utility Prisoners’ Dilemmas. In subsequent manuscripts we will attempt to generalize the existence mappings between PPD and UPD and demonstrate the practical implications of the non-existence results. In addition, formal characterization of conditions of under which conflict inducing growth may arise remains a high priority.

Bibliography

Brennan, G. 1973. “Pareto Desirable Redistribution in the Case of Malice and Envy,” *Journal of Public Economics* 2, 173-83.

Chaudhuri, A. 1985. “Formal Properties of Interpersonal Envy,” *Theory and Decision* 18, 301-12.

Hammond, P. 1987. “Envy,” in Eatwell, Milgate, and Newman eds, *The New Palgrarve Dictionary of Economics.* New York: The Stockton Press.

Rapoport, Anotol. 1987. “Prisoner’s Dilemma,” in Eatwell, Milgate, and Newman eds, *The New Palgrarve Dictionary of Economics.* New York: The Stockton Press.

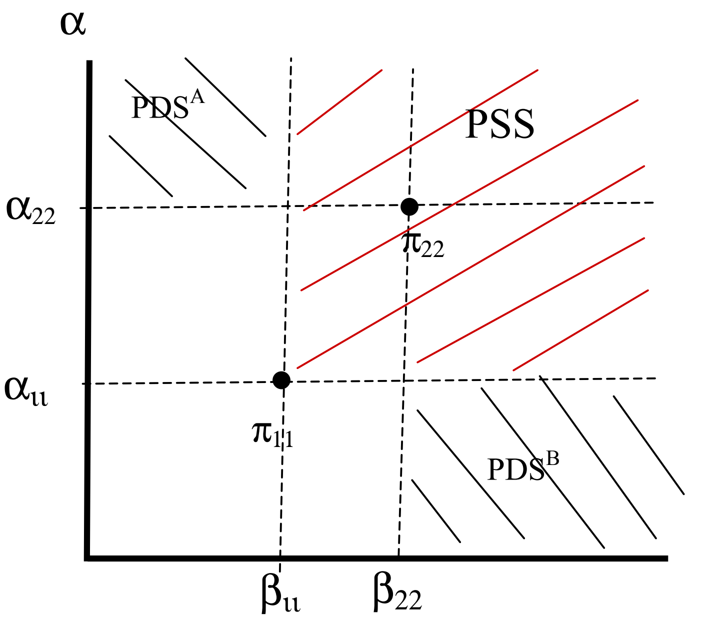

Figure 1a

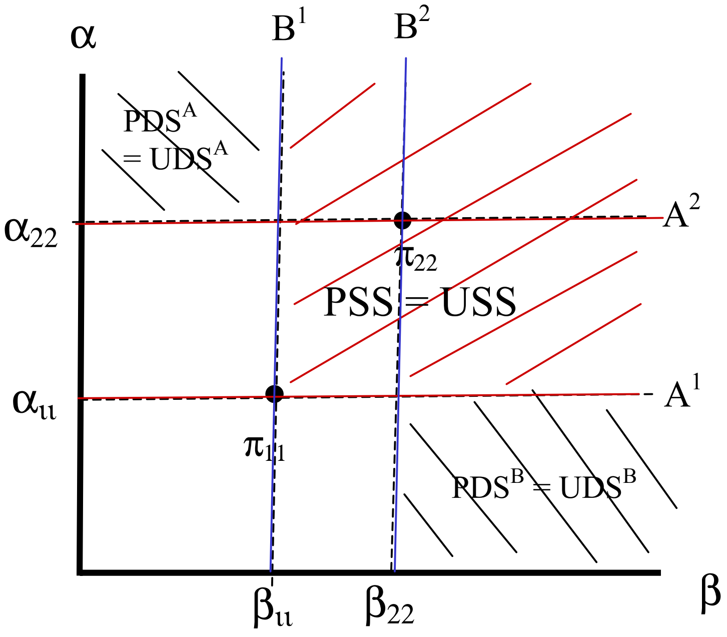

Figure 1b

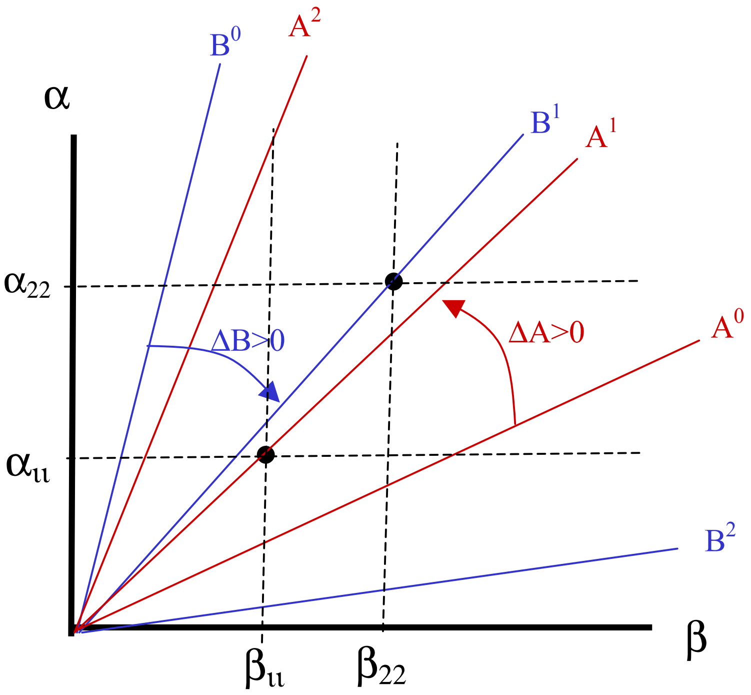

Figure 2

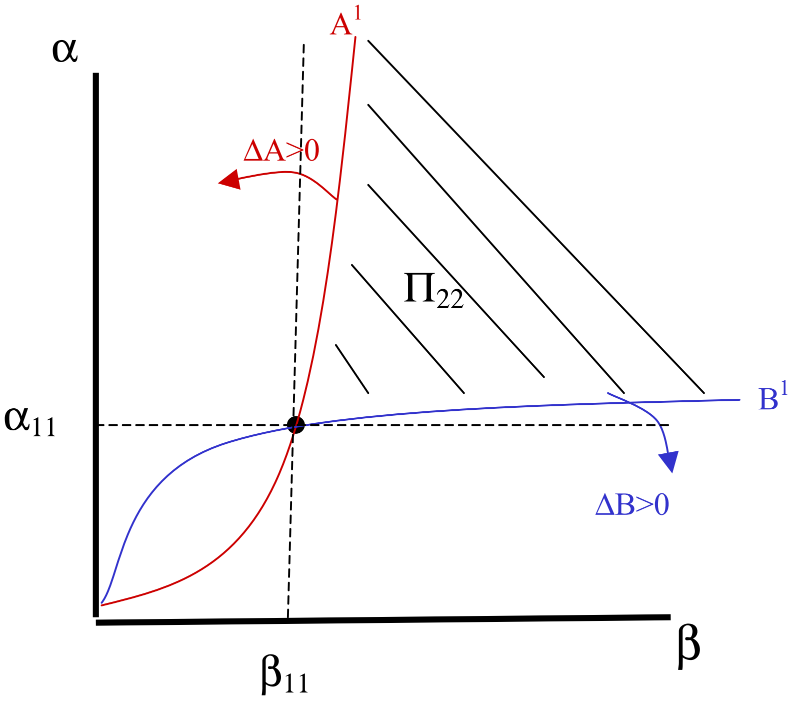

Figure 3

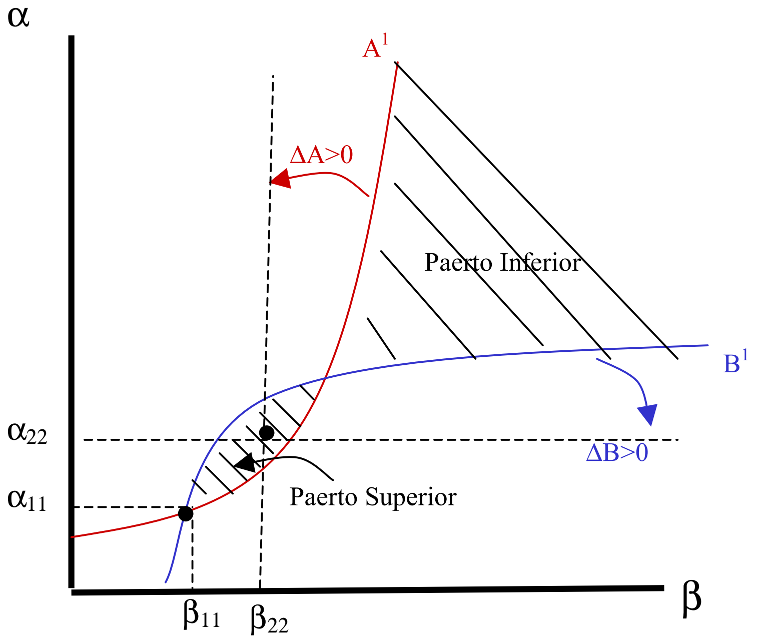
α
A^1
$\Delta A>0$
Paerto Inferior
B^1
α_{22}
$\Delta B>0$
α_{11}
Paerto Superior
β
β_{11}
β_{22}

Figure 4

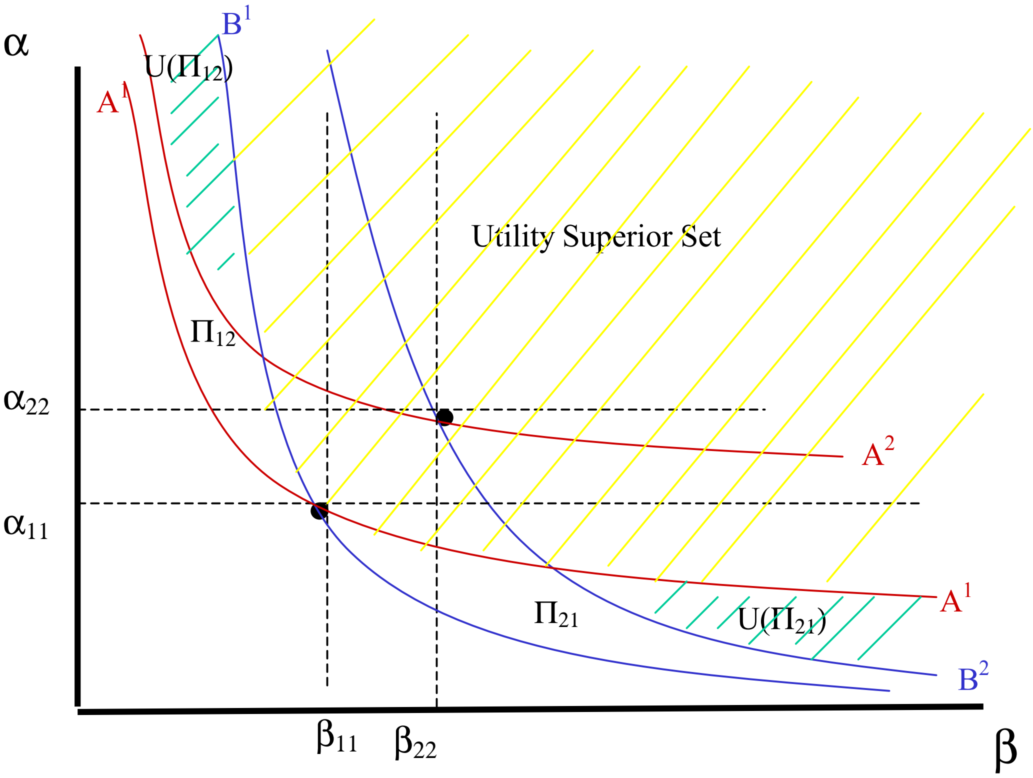
α
B^1
U(Π$_{12}$)
A^1
Utility Superior Set
Π$_{12}$
α$_{22}$
A^2
α$_{11}$
A^1
Π$_{21}$
U(Π$_{21}$)
B^2
β$_{11}$
β$_{22}$
β

Figure 5