# *RATIONAL COOPERATION*

# *IN ONE-SHOT SIMULTANEOUS PD-SITUATIONS*


***Ronald A. Heiner***

Center for Study of Public Choice

George Mason University

Fairfax, VA 22030

UNITED STATES

Fax: 703 993 2323

***Dieter Schmidtchen***

Universität des Saarlandes

Bldg. 31, Postfach 151150

D-66041 Saarbrücken

GERMANY

Fax: 49 681 302 359


***June 10, 1995***

# RATIONAL COOPERATION IN ONE-SHOT SIMULTANEOUS PD-SITUATIONS[1]

PART 1: Nash Equilibria for Symmetric Games With Fixed Forecast Probabilities

*"Game theory emanates from games such as chess or poker. Everyone knows that in these games players have to think ahead and devise a strategy based on expected countermoves from the other player. Such strategic interaction also characterizes many economic situations".*

1994 Nobel citation for game theorists (Nash, Harsanyi, Selten)

## I. INTRODUCTION

Given, as suggested in the Nobel citation, that players are going to base their strategic decisions on their expectations of each other's actions, they have an interest in avoiding mistaken expectations. We thus might want to consider the possibility that players are more likely to form expectations of each other's actions when those expectations are correct rather than mistaken. However, game theory has traditionally assumed players' expectations of each other's actions are statistically independent of whether such expectations will be confirmed or falsified by each other's actual behavior.

Suppose we thus consider how to model strategic interaction when guided by expectations that are statistically related to whether they will end up confirmed or falsified by the eventually revealed actions of other players. The following discussion presents a theory motivated in part by strategic relationships in "prisoner's dilemma" type situations (hereafter called PD-situations). To facilitate exposition, the two players in a PD-situation are hereafter referred to as "Adam" and "Eve"; so that we can easily distinguish the two players by their pronouns, "he" and "she", or "his" and "her", and so on.

A well known intuition for such situations is that since Adam would do relatively better for himself by defecting regardless of what Eve might do, it doesn't make any difference what he expects she will do. Even if he could forecast Eve's choice very accurately (perhaps even perfectly) Adam would still do better for himself by defecting rather than cooperating. Since this same inference applies to Eve, then the only result consistent with rationally self-interested players is for both Adam and Eve to always defect on each other regardless of what their forecasts of each other's choices might be. In short, rationally self-interested players will be better off always ignoring any expectations they may have about each other's potential cooperation, even though they both know that they would both be better off from mutually cooperating rather than always defecting.

Despite the above intuition, we will show that being able to forecast another player's actual cooperation better than pure chance can change Adam and Eves' strategic incentives in a one-shot simultaneous PD-situation. In particular, we will show that if they both have such ability (to forecast each other's actual choices better than pure chance), then "conditionally cooperative" Nash equilibria may also exist in addition to the traditional "always defect" equilibrium. By "conditionally cooperative" we mean Adam and Eve do not cooperate because of a behavioral "disposition" to act cooperatively, but rather they *selectively* cooperate *contingent* on their forecasting each other will also cooperate (cooperate if and only if

[1] Discussions in Saarbrucken, Germany during May-June 1994 and Fairfax, Virginia during September-October 1994 led one of us (RAH) to combine previous work in imperfect choice theory with noncooperative game theory to derive the main theoretical implications described below. The present paper is the first article resulting from our collaboration. We would like to thank the Volkswagen Foundation for financial support. In addition, we wish to thank the following persons for discussion and comments: Kenneth Arrow, Robert Axelrod, Ken Binmore, James Buchanan, Jürgen Eichberger, Robert Frank, Hartmut Kliemt, Christian Koboldt, Mathias Leder, Doug North, Joe Oppenheimer, Schmidt-Mohr, Andrew Schotter, Vernon Smith, Ulrich Witt. The usual disclaimer applies.

each other is forecasted to cooperate). Adam and Eve thereby "*contingently respond*" to their forecasts of each other, rather than always defecting regardless of their forecasts. Moreover, such contingent responding is motivated solely in order to maximize Adam's own expected payoff, given the way Eve responds to her forecast (including the option of ignoring her forecast); and vice versa concerning Eve's motivation for contingently responding to her forecasts of Adam's potential cooperation.

A key reason why conditionally cooperative equilibria have not been analyzed before is that traditional analysis assumes Eve believes the likelihood of Adam's cooperation, whatever it might be, is the same no matter what she chooses. Suppose, however, Adam can forecast Eve's *actual* choice better than pure chance (so that he is more likely to forecast Eve's cooperation when she actually cooperates rather than defects), and contingently responds to his forecast if and only if he forecasts Eve will also cooperate (thereby cooperating exactly as often as he forecasts Eve's cooperation). Doing so implies Adam is more likely to cooperate when Eve actually cooperates rather than defects; which thereby implies Eve's expected payoff may ***drop*** if she actually defects relative to that acheivable from actually cooperating. A similar implication applies to Adam's expected payoff if Eve can forecast his actual cooperation better than pure chance, and she also contingently responds to her forecast. Thus, it may be optimal for both Adam and Eve to contingently respond to their forecasts (rather than always defecting), given each other does so. This means contingent responding may be a "self-enforcing" Nash equilibrium, if Adam and Eve can both forecast each other's actual choices better than pure chance.

Another reason why conditionally cooperative equilibria have not been analyzed before is that traditional analysis assumes Adam and Eve use "exogenous messages" to forecast each other's potential cooperation; meaning the correlation between their individual messages exists independently of how Adam or Eve might respond to them. Such exogenous messages might be uncorrelated with each other (for example, when players observe independently flipped coins), or they might be correlated with each other (for example, when players observe temperatures at nearby locations).

Suppose, however, Adam and Eve can forecast each other's actual cooperation better than pure chance, but Eve decides to ignore her forecast (by always defecting regardless of her forecast) while Adam decides to contingently respond to his forecast (by cooperating if and only if he forecasts her cooperation). Eve's response strategy implies Adam is less likely to forecast her cooperation (because she always defects regardless of her forecast, and he is less likely to forecast her cooperation when she actually defects rather than cooperates); which in turn implies Adam is more likely to defect (because he contingently cooperates only if he forecasts her cooperation). This further implies Eve is also less likely to forecast Adam's cooperation (because he is more likely to defect, and she is also less likely to forecast his cooperation when he actually defects rather than cooperates). Thus, the likelihood of either Adam or Eve forecasting each other's cooperation itself depends on both of their strategies about whether to ignore or contingently respond to their individual expectations, or "forecast-messages", about each other's potential cooperation. That is, the likelihood of receiving their individual forecast-messages *"endogenously*" depends on how both players decide to respond to them (instead of being exogenous to their response strategies, as implied for coin flips or temperature observations).

The preceding implications mean Adam and Eve can forecast each other's actual cooperation better than pure chance *only with messages endogenously correlated with how they actually respond to them*. That is, what matters to forecasting better than pure chance is ***not*** the correlation between players'messages per se, but instead whether their forecast-messages are correlated with how they will actually respond to

them.[2] However, as noted above, traditional analysis assumes players use only exogenous information sources; thereby never investigating the strategic incentives implied by players endogenously forecasting each other's actual choices better than pure chance.

In what follows, we will analyze the above two themes; namely, the strategic implications of players forecasting each other's actual choices better than pure chance, and the necessity of endogenous messages required to achieve such forecasting ability. In order to focus on the essential logic as simply as possible, this paper analyzes only "symmetric" PD-situations in which both players' basic payoffs, strategy sets (about ignoring or contingently responding to their forecasts), and forecast-probabilities are the same. The sequel to this paper[3] analyses the general case for any PD-situation, whether symmetric or not; also allowing players' to "endogenously" vary their forecast-probabilities to maximize their individual expected payoffs.

Within the scope of symmetric PD-situations, this paper proceeds as follows. Section II presents the main definitions and theorems characterizing when *contingently-responding* (to players' forecasts about each other) is a Nash equilibrium in addition to *always-defecting* (where players always ignore their forecasts). Section III derives explicit formulas for calculating numerical examples to illustrate the formal analysis. Section IV calculates three examples. The third example illustrates that conditionally cooperating can be a Nash equilibrium no matter how close players' forecasting ability gets to the limit of forecasting no better than pure chance. Section V further analyses the intuition behind the analysis, similar to that described above. Section VI formally links the analysis to the above distinction between endogenous versus exogenous information sources. Section VII compares the resulting analysis with traditional game theory models allowing either "mixed strategies" or "correlated equilibria". Section VIII discusses more general issues about strategically interdependent choices creating incentives toward statistically interdependent behavior; plus a further discussion of why always defecting is *not* implied by "revealed preference" theory. Section IX briefly describes how conditionally cooperative equilibria imply a behavioral sensitivity to changes in the relative differences between players' PD-payoffs; in general agreement with numerous previous experiments. Section X outlines companion papers which further develop related theory and experimental topics. Section XI concludes with a brief discussion about the historical significance of cooperation in one-shot PD-situations.

## II. DEFINITIONS & NASH EQUILIBRIA

DEFINITION 1 (Prisoner's Dilemma or PD-Situations)

> Four payoff levels arise in prisoner's dilemma or PD-type situations: R, the "reward" from both players mutually cooperating; $P < R$, the relative "penalty" from both players mutually defecting on each other; $T > R$, the "temptation payoff" from unilaterally defecting when the other player cooperates; $S < P$, the "sucker's payoff" from unilaterally cooperating when the other player defects. These inequalities are equivalent to the following ordinal PD-payoff ranking: $T > R > P > S$.

DEFINITION 2 (Conditional Forecast Probabilities)

> 1. Let Adam's actual choice to cooperate or defect be denoted by $C_A$ and $D_A$ respectively; and similarly for Eve's actual choices, denoted $C_E$ and $D_E$. Also let $C_A^f$ and $D_A^f$ denote that Adam

[2] In particular, such messages must be correlated with those "states of mind" which lead players to make actual choices; or with whatever behavioral mechanisms actually generate their choices (as discussed in Section VI below).

[3] For an outline of this paper, see Section X below.

forecasts Eve will actually choose $C_E$ and $D_E$ respectively; and similarly for Eve's forecasts of Adam's actual choice, denoted $C_E^f$ and $D_E^f$.

2. Adam's conditional probability of "rightly forecasting" $C_A^f$ when Eve actually chooses $C_E$ is denoted, $r_A = p(C_A^f \mid C_E)$. Conversely, his conditional probability of "wrongly forecasting" $C_A^f$ when Eve actually chooses $D_E$ is denoted, $w_A = p(C_A^f \mid D_E)$. Analogous definitions apply to Eve's conditional chance of either rightly or wrongly forecasting Adam's actual cooperation, denoted $r_E = p(C_E^f \mid C_A)$ and $w_E = p(C_E^f \mid D_A)$ respectively. The (r, w) probabilities are written without subscripts when they may refer to either player's forecasts.

DEFINITION 3 (Positive versus Zero Forecasting Ability)

Adam can forecast Eve will actually choose $C_E$ "*better than pure chance*" if and only if he is more likely to forecast her cooperation when she actually chooses $C_E$ than when she does not; so that $r_A > w_A$ is possible. Adam is then said to have "*positive forecasting ability*". Otherwise, Adam can forecast "*no better than pure chance*", and is said to have "*zero forecasting ability*"; if $r_A \equiv w_A$ is the only possibility. Analogous definitions apply to Eve's forecasting ability.

COMMENTS:

1. Consider an experimental setting where Adam and Eve are separated from time $t_1$ to $t_2$; during which they must each choose privately with no communication between them. At a later time $t_3$ their actual choices are revealed to each other. Let D denote the time interval from $t_1$ to $t_2$. Adam and Eve may have had opportunity to communicate before D, but whatever they may have said cannot "bind" them to any particular choice once they are physically separated during D. Moreover, because they are separated, neither player knows who actually chose first or second during D. Nevertheless, they can still attempt to forecast (while separated during D) what each other's eventually revealed choice will turn out to be at $t_3$; where $t_3$ necessarily occurs after D is completed.

2. In the above setting, forecasting better than pure chance ($r_A > w_A$) means Adam is more likely to forecast Eve's cooperation when her eventually revealed choice will confirm rather than falsify his forecast (because she actually chose to cooperate rather than defect during D).

3. Forecasting no better than pure chance means Adam's likelihood of forecasting Eve's cooperation is independent of whether her eventually revealed choice will confirm or falsify his forecast. For example, suppose Adam rolls a six-sided die to forecast Eve's revealed choice at $t_3$; and forecasts her cooperation or defection if the number (1, 2, 3, 4) or (5, 6) respectively faces upward when the die stops moving. Then the likelihood of Adam forecasting Eve's cooperation is 2/3 regardless of what her eventually revealed choice turns out to be; so that $(r_A, w_A) = (2/3, 2/3)$.

4. Recent experiments by Robert Frank, et. al. asked players to both choose between C or D, and forecast the action chosen by the other player during a separation interval D like that discussed above. The subjects in these experiments were able to forecast each other's actual cooperation better than pure chance in one-shot simultaneous PD-situations with explicit money payoffs[4]. For example, the conditional forecast

4 See chapter 7 of *Passions Within Reason*, W. W. Norton, 1991, by Frank; and "The Evolution of One-Shot Cooperation: An Experiment," *Ethology and Sociobiology*, 14, 1993, pp. 247-256, by Frank, Gilovich, & Regan.

probabilities generated in two different experiments were (r, w) = (.88, .62) and (r, w) = (.89, .59); resulting from sample sizes of 122 and 198 respectively. We thus have direct experimental evidence that positive forecasting ability is possible in empirical situations with PD-payoffs (T > R > P > S).

DEFINITION 4 (Response Strategies: Conditionally Cooperating versus Always Defecting)

1. A "response-strategy" is an ordered pair $XY_A$; where X is Adam's actual choice when he forecasts Eve will choose $C_E$, and Y is his actual choice when he forecasts Eve will choose $D_E$. The subscript A thus refers to Adam's actual responses to forecasting Eve will either cooperate or defect respectively. The ordered pair $XY_E$ similarly represents Eve's actual choices in response to forecasting Adam will either cooperate or defect respectively.

2. An ordered pair $XY_A$ or $XY_E$ thus reflects the strategic perspective suggested in the Nobel citation quoted at the beginning; namely, "everyone knows that ... players have to think ahead and devise a strategy based on expected countermoves from the other player." Such ordered pairs are thus called "*response strategies*". The A and E subscripts are deleted when the response strategy XY may refer to either player.

2. The response strategy DD means to "always defect" no matter what the other player is forecasted to do, and CD means to "conditionally cooperate" if and only if the other player is also forecasted to cooperate.

COMMENTS:

1. The strategy combination ($DD_A$, $DD_E$) represents the traditional "dominant strategy" solution for simultaneous, single-shot PD-situations; where both players always defect no matter what they might forecast about each other's choice. ($DD_A$, $DD_E$) thus implies both players have zero chance of actually cooperating.

2. On the other hand, the strategy combination ($CD_A$, $CD_E$) does **not** imply any general "disposition" or "tendency" for players to cooperate; but instead represents a "conditional strategy" to potentially cooperate contingent on what each player forecasts the other player will actually do. Consequently, ($CD_A$, $CD_E$) does not imply either player will necessarily cooperate or not. Instead, it implies positive probabilities of either zero, one, or both players actually cooperating.

3. A pair of response strategies ($XY_A$, $XY_E$) might be interpreted as a possible "agreement" in which Adam and Eve exchange promises to respond to their forecasts of each other's choice according to $XY_A$ and $XY_E$ respectively. By this we do not necessarily mean Adam and Eve have actually met and discussed such an agreement. Rather, their behavior is consistent with a "hypothetical agreement" having been made.

However, care should be taken to avoid this interpretation, because the term "hypothetical agreement" does **not** mean there is any "moral obligation" or some other cost associated with violating promises, thereby causing players to act potentially contrary to their self-interest in the absence of such an agreement. We are instead dealing with *rationally self-interested players* who never keep agreements "because" they promised to do so. Rather, they make promises only when it is already in their self-interest to keep them in the first place. Such players will thus give credence only to truly "self-enforcing" agreements

which do not require any potential "commitment device" or "enforcement mechanism" beyond their own self-interest in order to motivate voluntary compliance.

4. Therefore, a strategy combination ($XY_A$, $XY_E$) may not involve any actual exchange of promises or any explicit agreement actually communicated between the players. It can nevertheless be used to represent the behavior of rational players who follow strategies only when doing so is truly "self-enforcing" in order to maximize their own expected payoff given each other's strategy. That is, no actual agreement, enforcement mechanism, or anything else is needed to explain such players' behavior beyond what their rational self-interest already implies. Subsequent discussion thus refers only to *self-enforcing strategies*; as defined next.

DEFINITION 5 (Self-Enforcing Strategies and Nash Equilibrium)

> A pair of response strategies ($XY_A$, $XY_E$) is "self-enforcing" if and only if it is optimal (expected payoff maximizing) for each player to actually follow its strategy conditional on the other player also following its strategy; so that neither player has a unilateral incentive to deviate from its response strategy. Such a strategy combination thus represents a Nash equilibrium.

As noted in the introduction, this paper focuses on "symmetric" PD-situations in order to develop the essential logic of the analysis as simply as possible. Symmetric situations mean players' strategy sets, basic payoffs, and forecast probabilities are the same for both Adam and Eve.

DEFINITION 6 (Symmetric PD-Situations)

> Let $S$ = {DD, CD} denote the above defined pair of strategies for responding to a player's forecast of the other player's actual choice to either cooperate or defect, and let $\xi$ = (T,R,P,S) denote a set of basic payoffs potentially received by a player. A "symmetric" one-shot PD-situation means both players have the same basic payoffs x and same strategy set $S$; as well as the same conditional forecast probabilities, denoted $(r_A, w_A) = (r_E, w_E) = (r, w)$. Such a symmetric one-shot simultaneous PD-game is denoted, PD($S$, $\xi$, r, w).

The next definition is introduced in order to characterize when particular strategy combinations from $S \times S$ represent self-enforcing Nash equilibria.[5]

DEFINITION 7 (Sufficiently Reliable Forecasts)

> A player's forecast of the other player's actual choice is **"sufficiently reliable"** if and only if,
>
> $$r(R - S) - w(T - P) \geq P - S;$$
>
> or equivalently, (1)

5 Besides {DD, CD} there are two other possible response strategies {CC, DC}; corresponding to always cooperating regardless of one's forecast, and defecting if and only if the other player is forecasted to cooperate (the opposite to strategy CD). Neither of these strategies can be Nash equilibria no matter what players' forecast probabilities might be. Consequently, they are omitted in order to focus on the essential logic in this paper. The sequel to this paper (Part II; see Section X below) explicitly includes all four response strategies; also allowing players to "endogenously" vary their forecast probabilities.

$$(r - w)[R - P] \geq (1 - r)[P - S] + w[T - R];$$

where R - P is the net gain from mutual cooperation over mutual defection; T - R is the net gain from unilaterally defecting (when the other player cooperates); and P - S is the net loss from unilaterally cooperating (when the other player defects).

DEFINITION 8 (Cooperation Line)

The set of (r, w) probabilities which exactly satisfy inequality (1) is called the **cooperation line**. (r, w) points on or above this line satisfy inequality (1); points below this line violate inequality (1).

NOTE:

1. Figure 1 shows a "unit probability box" with a cooperation line implied from inequality (1); along with the formulas for the top and left intercepts of the line. Notice that players' basic payoffs x = (T, R, P, S) uniquely determine the slope and position of the line [because the intercept formulas depend only on these payoffs].

**FIGURE 1 ABOUT HERE**

2. With inequality (1) and its corresponding cooperation line defined, we can now describe the possible self-enforcing strategies for any symmetric one-shot simultaneous game, PD($S$,$\xi$,r,w).

THEOREM 1 (Self-Enforcing, Nash-Equilibrium Strategies)

The following statements characterize when potential strategy combinations from $S \times S$ represent self-enforcing Nash equilibria for any symmetric one-shot simultaneous game, PD($S$, $\xi$, r, w).

1. If players' forecasts of each other's actual choice are not sufficiently reliable [meaning (r, w) violates inequality (1), so that (r, w) lies below the cooperation line], then ($DD_A$, $DD_E$) is the unique, dominant-strategy Nash equilibrium. Hence, ($DD_A$, $DD_E$) is the only possible self-enforcing Nash equilibrium.

2. If players' forecasts are sufficiently reliable [meaning (r, w) satisfies inequality (1), so that (r, w) lies on or above the cooperation line], then ($DD_A$, $DD_E$) and ($CD_A$, $CD_E$) are **both** Nash equilibria. Hence, both ($DD_A$, $DD_E$) and ($CD_A$, $CD_E$) are self-enforcing strategy combinations.

COMMENTS (Theorem 1 is proven in the Appendix):

1. Frank's experiments (1991, 1993) directly support Theorem 1. In particular, subjects in his experiments did not cooperate due to a stable "personality type" or "disposition"; as argued by Gauthier, Howard, and others. Rather, they cooperated when they forecasted the other player would also cooperate. Frank's summary of experimental results is instructive (1991; pages 141-142):

> "these findings ... should not be interpreted as evidence of stable personality types called cooperators and defectors. On the contrary, we found that at least some of our subjects did not consistently follow either strategy: 13 of them (21 percent) cooperated with one of their partners but not with the other. ... A pattern

observed in all three versions of the experiment ... was for subjects to behave in the same way they predicted their partners would. In the basic version, for example, 83 percent of the subjects who predicted their partners would cooperate also cooperated themselves. Similarly, 85 percent of the subjects who predicted defection also defected themselves."

2. The intercept formulas for the cooperation line in Figure (1) imply it can be shifted arbitrarily close to the $45^0$-line corresponding to zero forecasting ability (where r = w) by reducing T - R and P - S sufficiently relative to R - P; while still preserving the ordinal payoff ranking, T>R>P>S. Consequently, forecasting better than pure chance (r > w) implies there exist payoffs for which "conditionally cooperating" according to ($CD_A$, $CD_E$) is a self-enforcing equilibrium even though the ordinal PD-payoff ranking (T > R > P > S) is still satisfied. That is, ($CD_A$, $CD_E$) can be a self-enforcing equilibrium no matter how close we get to the limit of zero forecasting ability. This implication is more precisely described in the following definition, theorem, and corollary.

DEFINITION 9 (Cost-Benefit Ratio for PD Cooperation)

1. Let x = max[T - R, P - S], and y = R - P. The variable x is a measure of the opportunity cost of cooperating; either from forsaking a potential net gain T - R by unilaterally defecting on the other player's cooperation; or from risking a potential net loss P - S by unilaterally cooperating when the other player defects. The variable y measures the potential benefit from mutually cooperating compared to mutually defecting.

2. x/y can thus be interpreted as a "cost-benefit ratio" which measures the potential costs from cooperating relative to the potential benefits from doing so, compared to otherwise mutually defecting.

THEOREM 2 (Forecasting Better Than Pure Chance Implies Inequality (1) Holds As $x/y \rightarrow 0$)

1. Forecasting better than pure chance (r > w) implies there exist PD-payoffs (T > R > P > S) such that inequality (1) is satisfied for sufficiently small but still positive x/y > 0; so that players' forecasts are guaranteed to be sufficiently reliable as the cost-benefit ratio $x/y \rightarrow 0$.

DEFINITION 10 (Pure Prisoner's Dilemma Games)

1. Let PD(*S*, r, w) denote the set of all symmetric one-shot simultaneous PD-games, PD(*S*,ξ,r,w); such that players' strategy sets and forecast probabilities equal respectively *S* and (r, w).

2. PD(*S*, r, w) is then called a "*pure prisoner's dilemma*" if and only if ($DD_A$, $DD_E$) is the only self-enforcing equilibrium for all PD[*S*, ξ = (T,R,P,S), r, w] that satisfy the ordinal payoff ranking, T > R > P > S.

COROLLARY 1 (Resolution of the "Paradox of PD-Cooperation")

PD(*S*, r, w) is a pure prisoner's dilemma if and only if neither player can forecast the other player's actual choice better than pure chance; so that r = w is satisfied. Otherwise, forecasting better than pure chance (r > w) guarantees there exist T > R > P > S such that ($CD_A$, $CD_E$) is a self-enforcing Nash equilibrium for the corresponding one-shot simultaneous game PD[*S*, x = (T,R,P,S), r, w].

COMMENT: Corollary 1 implies the ordinal payoff ranking traditionally associated with PD-situations (T > R > P > S) is **not** sufficient to guarantee defection is a "dominant strategy". That is, ($DD_A$,$DD_E$) is **not** necessarily the unique dominant equilibrium; so that ($DD_A$, $DD_E$) may **not** be the only possible self-enforcing equilibrium between the players. Rather, ($CD_A$, $CD_E$) may also be a self-enforcing equilibrium so long as players can forecast each other's actual choice better than pure chance. Consequently, any positive forecasting ability better than pure chance implies a whole range of one-shot PD-situations are now compatible with "rationally cooperating"; thereby resolving the seeming "paradox" of rational behavior never allowing cooperation in such situations.

## III. STATISTICAL FORMULAS FOR CALCULATING EXAMPLES

In order to calculate explicit examples of Theorem 1, we must first explain how to calculate "joint choice-probabilities" of Adam and Eves' actual choices to either cooperate or defect.

DEFINITION 11 (Joint-Probabilities of Players' Actual Choice Combinations)

> The "*joint choice-probabilities*" of both players actually cooperating, only one of them actually cooperating, or both players actually defecting are denoted respectively; $p(C_A, C_E)$, $p(C_A, D_E)$, $p(D_A, C_E)$, $p(D_A, D_E)$.

These joint-probabilities determine the likelihood of Adam and Eve actually receiving each of their four PD-payoffs (T, R, P, S). Their expected payoff formulas are denoted respectively,

$$U_A = Rp(C_A, C_E) + Sp(C_A, D_E) + Tp(D_A, C_E) + Pp(D_A, D_E) \qquad (2a)$$

$$U_E = Rp(C_A, C_E) + Tp(C_A, D_E) + Sp(D_A, C_E) + Pp(D_A, D_E) \qquad (2b)$$

DEFINITION 12 (Players' Individual Choice Probabilities)

> The following conditional and unconditional "*choice probabilities*" are defined for Adam: $u_A = p(C_A \mid C_E)$, $1 - u_A = p(D_A \mid C_E)$, $v_A = p(C_A \mid D_E)$, $1 - v_A = p(D_A \mid D_E)$; and $z_A = p(C_A)$, $1 - z_A = p(D_A)$. Analogous definitions apply to Eve's choice probabilities: $u_E = p(C_E \mid C_A)$, $1 - u_E = p(D_E \mid C_A)$, $v_E = p(C_E \mid D_A)$, $1 - v_E = p(D_E \mid D_A)$; and $z_E = p(C_E)$, $1 - z_E = p(D_E)$. Players' conditional choice probabilities (which depend on each other's choices) are thus represented with the letters u and v, and their unconditional choice probabilities are represented with the letter z. Subscripts are deleted when (u, v, z) may refer to either Adam or Eve's choice probabilities.

COMMENTS:

1. The reason for defining both conditional and unconditional choice probabilities is that positive forecasting ability (r > w) implies that players' individual choices may not be statistically independent of each other, if either player conditionally responds to its forecast according to CD. Consequently, the joint-probabilities used in equations (2a,b) cannot necessarily be calculated by multiplying players' unconditional choice probabilities, $z_A$ and $z_E$. For example, $p(C_A, C_E)$ can be expressed as either of the following two

multiples: $p(C_A \mid C_E)p(C_E) = u_A z_E$, or $p(C_E \mid C_A)p(C_A) = u_E z_A$. These two multiples do not simplify to the single multiple $z_A z_E$ except for statistically independent choices, such that $z_A = u_A$ and $z_E = u_E$.

2. Each of the joint-probabilities in equations (2a,b) can similarly be expressed as two probability multiples [by reversing the order of conditional probabilities used to calculate each joint-probability]:

$$p(C_A, C_E) = u_A z_E = u_E z_A \tag{3a}$$

$$p(D_A, C_E) = (1 - u_A) z_E = v_E(1 - z_A) \tag{3b}$$

$$p(C_A, D_E) = v_A(1 - z_E) = (1 - u_E) z_A \tag{3c}$$

$$p(D_A, D_E) = (1 - v_A)(1 - z_E) = (1 - v_E)(1 - z_A) \tag{3d}$$

3. Adding equations (3a) and (3b) implies, $z_E = u_E z_A + v_E(1 - z_A)$. Similarly adding equations (3c) and (3d) (plus rearranging terms) also implies, $z_A = u_A z_E + v_A(1 - z_E)$. We thus have the following two equations for players' unconditional choice probabilities:

$$z_A = u_A z_E + v_A(1 - z_E) \qquad \& \qquad z_E = u_E z_A + v_E(1 - z_A) \tag{4a,b}$$

Equations (4a,b) imply that players' unconditional choice probabilities ($z_A$, $z_E$) are statistically interdependent with each other and with both of their conditional choice probabilities ($u_A$, $v_A$, $u_E$, $v_E$).

4. Adam's unconditional chance of cooperating thereby depends on his conditional chance of cooperating when Eve cooperates or not, which depends on how her unconditional chance of cooperating depends on her conditional chance of cooperating when Adam cooperates or not, which in turn depends on Adams's unconditional chance of cooperating, and so on; leading to a statistical "infinite regress". For example, suppose we interpret the statistical relationships of equations (4a,b) as players "thinking" about responding to each other. Such hypothetical responding leads to an infinite regress whereby Adam considers his likely response knowing Eve is also considering how to respond to his likely response, and knowing Eve knows that he knows she is responding to his likely response, and so on.

5. Despite the statistical interdependence and infinite regress just noted, we can still obtain determinant values for both players' unconditional chances of cooperating or not. The reason is that equations (4a,b) are linear in all their variables; so that we can solve them simultaneously for ($z_A$, $z_E$) as functions of the remaining variables ($u_A$, $v_A$, $u_E$, $v_E$). The resulting formulas give a determinant statistical prediction of both players' unconditional likelihood of cooperating or not:

$$z_A = \frac{u_A v_E + v_A(1 - v_E)}{1 - (u_A - v_A)(u_E - v_E)} \qquad \& \qquad z_E = \frac{u_E v_A + v_E(1 - v_A)}{1 - (u_A - v_A)(u_E - v_E)} \tag{5a,b}$$

Notice that no infinite regress to ever higher "meta-levels" of hypothetical responding arises in equations (5a,b). Instead, there is only "one-round" of mutual statistical interdependence between players' actual choices; which leads directly to determinant calculations of the probabilities of these choices.

6. By substituting (5a,b) into (3a,b,c,d), players' joint probabilities can be written as functions of their conditional choice probabilities ($u_A$, $v_A$, $u_E$, $v_E$). This in turn implies that players' expected payoff formulas (2a,b) can also be expressed as functions of these same variables, plus their PD-payoffs $\xi = (T,R,P,S)$. These two expected payoff functions are written:

$$U_A = \pi^A[(u_A, v_A), (u_E, v_E); \xi] \quad \& \quad U_E = \pi^E[(u_A, v_A), (u_E, v_E); \xi] \tag{6a,b}$$

7. Next consider how players' conditional forecast and conditional choice probabilities are related. Recall that response strategy DD implies a player always defects or never cooperates regardless of its forecast of the other player's choice; which implies it also never cooperates regardless of the other player's actual choice. We thus have the following implication,

DD implies (u, v) = (0, 0); for either player's conditional choice probabilities (7a)

On the other hand, response strategy CD implies that a player cooperates exactly as often as it forecasts the other player will cooperate [because CD means it cooperates if and only if it forecasts the other player will also cooperate]. We thus have the implication,

CD implies (u, v) = (r, w); for either player's conditional choice probabilities (7b)

8. By substituting (7a) or (7b) into (5a,b) and then substituting into (3a,b,c,d), the players' four joint-probabilities [$p(C_A, C_E)$, $p(D_A, C_E)$, $p(C_A, D_E)$, $p(D_A, D_E)$] become functions of their conditional forecast probabilities, ($r_A$, $w_A$) and ($r_E$, $w_E$), plus their pair of response strategies ($XY_A$, $XY_E$); where $XY_A$ and $XY_E$ respectively denote Adam and Eves' response strategies from $S = \{DD, CD\}$. Each player's expected payoff formula (2a,b) thereby also becomes a function of both players' conditional forecast probabilities and both of their response strategies; plus their basic payoffs x = (T,R,P,S). We can thus re-express players' expected payoff functions (6a,b) as follows,

$$U_A = \pi^A[(r_A, w_A), (r_E, w_E); (XY_A, XY_E); \xi] \tag{8a}$$

$$U_E = \pi^E[(r_A, w_A), (r_E, w_E); (XY_A, XY_E); \xi] \tag{8b}$$

Note the above substitutions [(7a) or (7b) into (5a,b)] also imply that formulas (5a,b) for players' unconditional choice probabilities ($z_A$, $z_E$) are functions of their forecast probabilities and their response strategies; denoted as follows,

$$z_A = z^A[(r_A, w_A), (r_E, w_E); (XY_A, XY_E)] \quad \& \quad z_E = z^E[(r_A, w_A), (r_E, w_E); (XY_A, XY_E)] \tag{8c,d}$$

9. Finally, recall that Theorem 1 applies to symmetric games where players' conditional forecast probabilities are equal. Expected payoffs functions (8a,b) can thus be simplified to,

$$U_A = \pi^A[(r, w); (XY_A, XY_E); \xi] \quad \& \quad U_E = \pi^E[(r, w); (XY_A, XY_E); \xi] \tag{9a,b}$$

## IV. EXAMPLES WITH ZERO AND POSITIVE FORECASTING ABILITY

### 1. Example With Zero Forecasting Ability

With the above equations and statistical relationships formally specified in (2a,b) - (9a,b), we can now calculate examples of Theorem 1. First consider an example with zero forecasting skill, corresponding to r = w; such as (r, w) = (.5, .5) for both players, and (T,R,P,S) = (10, 8, 6, 4) for both players. It is easy to see that (.5,.5) lies below the cooperation line implied by (10, 8, 6, 4) [because (.5)(8 - 4) - (.5)(10 - 6) = 0 < (6 - 4) = 2]; so that inequality (1) is violated. Thus, ($DD_A$,$DD_E$) is the unique dominant equilibrium according to Part 1 of Theorem 1. Let us show how to verify this result for the one-shot simultaneous game PD[*S*,(10,8,6,4),.5,.5].

1. First calculate players' joint-probabilities implied from the strategy pair ($DD_A$, $DD_E$). Substituting implication (7a) into equations (5a,b) implies, ($z_A$, $z_E$) = (0, 0);. Further substituting into equations (3a,b,c,d) implies that players' joint-probabilities all equal 0 except for the chance of them both defecting which equals 1. That is, [p($C_A$, $C_E$), p($D_A$, $C_E$), p($C_A$, $D_E$), p($D_A$, $D_E$)] = (0, 0, 0, 1). This is of course what one would expect, since ($DD_A$, $DD_E$) implies both players always defect (hence never cooperate) regardless of their forecasts.

2. Next calculate the joint-probabilities implied from the strategy pair ($CD_A$, $CD_E$). Substituting (7b) into (5a,b) implies,

$$z_A = \frac{(.5)(.5) + (.5)(1 - .5)}{1 - (.5 - .5)^2} = \frac{.25 + .25}{1} = .5 = z_E \qquad (10)$$

Thus, ($z_A$, $z_E$) = (.5, .5); which further implies by substituting into (3a,b,c,d), that players' joint-probabilities satisfy, [p($C_A$, $C_E$), p($D_A$, $C_E$), p($C_A$, $D_E$), p($D_A$, $D_E$)] = (.25, .25, .25, .25).

3. Next calculate the joint-probabilities implied from the strategy pair ($CD_A$, $DD_E$). Calculations like those in steps 1 and 2 above imply; ($z_A$, $z_E$) = (.5, 0), and [p($C_A$, $C_E$), p($D_A$, $C_E$), p($C_A$, $D_E$), p($D_A$, $D_E$)] = (0, 0, .5, .5). Note that either (7a) or (7b) must be used in substituting into (5a,b) and (3a,b,c,d), depending on which player uses strategy DD or CD respectively. Similar calculations also imply the latter joint-probabilities equal respectively (0, .5, 0, .5), for the strategy pair ($DD_A$, $CD_E$).

4. With the preceding three steps finished, we can calculate players' expected payoffs according to formulas (2a,b). For example, strategy pair ($CD_A$, $CD_E$) implies,

$$U_A = \pi^A[(.5, .5); (CD_A, CD_E); (10, 8, 6, 4)] = 10(.25) + 8(.25) + 6(.25) + 4(.25) = 7 \qquad (11a)$$

$$U_E = \pi^E[(.5, .5); (CD_A, CD_E); (10, 8, 6, 4)] = 10(.25) + 6(.25) + 8(.25) + 4(.25) = 7 \qquad (11b)$$

Table 1 summarizes the above joint-probability and expected payoff calculations implied from different combinations of strategies from *S* = {DD, CD}. The strategic implications of Table 1 can be represented with a traditional "normal" form *strategy matrix*. The numbers in the matrix are the expected

payoffs ($U_A$, $U_E$) shown in Table 1, which result from players choosing different combinations of strategies CD or DD. Note how the strategy matrix implies ($DD_A$, $DD_E$) is the unique dominant strategy equilibrium; as also implied by Part 1 of Theorem 1. This equilibrium implies a degenerate joint-probability distribution (0, 0, 0, 1); shown in Table 1.

**TABLE 1 & ITS STRATEGY MATRIX ABOUT HERE**

2. Example With Positive Forecasting Ability

Consider the same example as before, except that now players have positive forecasting skill, such as (r, w) = (.9, .3) for both players. In this situation, (.9, .3) lies above the cooperation line [because (.9)(8 - 4) - (.3)(10 - 6) = 2.4 > (6 - 4) = 2]; so that inequality (1) is satisfied. Thus, Part 2 of Theorem 1 implies ($DD_A$, $DD_E$) and ($CD_A$, $CD_E$) are both self-enforcing Nash equilibria for the one-shot simultaneous game, PD[*S*, (10, 8, 6, 4), .9, .3]. Let us calculate as in the prior example to verify this result.

1. First calculate ($z_A$, $z_E$) for the strategy pair ($CD_A$, $CD_E$) by substituting (7b) into (5a,b), to obtain:

$$z_A = \frac{(.9)(.3) + (.3)(1 - .3)}{1 - (.9 - .3)^2} = \frac{.27 - .21}{1 - (.6)^2} = \frac{.48}{.64} = .75 = z_E \qquad (12)$$

Thus, ($z_A$, $z_E$) = (.75, .75); which further implies by substituting into (3a,b,c,d), that players' joint-probabilities satisfy, [$p(C_A, C_E)$, $p(D_A, C_E)$, $p(C_A, D_E)$, $p(D_A, D_E)$] = (.675, .075, .075, .175). These probabilities are calculated by substituting (7b) into (5a,b), and then substituting into (3a,b,c,d), to obtain:

$$p(C_A, C_E) = z_A r_E = z_E r_A = (.75)(.9) = .675 \qquad (13a)$$

$$p(D_A, C_E) = (1 - r_A) z_E = (1 - .9)(.75) = w_E(1 - z_A) = (.3)(1 - .75) = .075 \qquad (13b)$$

$$p(C_A, D_E) = z_A(1 - r_E) = (.75)(1 - .9) = (1 - z_E) w_A = (1 - .75)(.3) = .075 \qquad (13c)$$

$$p(D_A, D_E) = (1 - z_A)(1 - w_E) = (1 - z_E)(1 - w_A) = (1 - .75)(1 - .3) = .175 \qquad (13d)$$

2. In like manner, players' joint-probabilities can be calculated for the other three strategy pairs {($CD_A$,$DD_E$), ($DD_A$, $CD_E$), ($DD_A$, $DD_E$)}; except that both (7a) and (7b) must be used depending on which player uses strategy DD or CD respectively. Table 2 summarizes these joint-probabilities and associated expected payoff calculations implied by different strategy combinations of either DD or CD. As in the first example, the strategic implications of Table 2 are represented in a normal form strategy matrix. The numbers in the matrix are the expected payoffs ($U_A$, $U_E$) resulting from players choosing different combinations of response strategies CD or DD.

**TABLE 2 & ITS STRATEGY MATRIX ABOUT HERE**

Note how the strategy-matrix implies there are two Nash equilibria, ($CD_A$, $CD_E$) and ($DD_A$, $DD_E$); in accordance with Part 2 of Theorem 1. Consequently, "always defecting" regardless of what players forecast about each other [corresponding to ($DD_A$, $DD_E$)] is not the only Nash equilibrium. Rather, each player "conditionally-cooperating" according to CD is an optimal strategic reaction to the other player also conditionally cooperating in the same manner; so that ($CD_A$, $CD_E$) becomes a self-enforcing equilibrium.

3. Example With A Small PD Cost-Benefit Ratio

The next example illustrates Theorem 2 and Corollary 1, about ($CD_A$, $CD_E$) being a self-enforcing equilibrium even with little positive forecasting skill above pure chance; provided the PD cost-benefit ratio x/y is sufficiently small but still positive [see Definition 9 above]. In particular, suppose players' basic payoffs are x = (25.2, 25, 5, 4.8); which implies, x = .2, y = 20, and x/y = .01. Suppose also players can forecast each others' actual choice only slightly better than pure chance; such as (r, w) = (.53, .5). This means each player has only 53 to 50 odds of rightly rather than wrongly forecasting the other player's actual cooperation. For example, Adam has 50-50 odds of mistakenly forecasting Eve's cooperation when her eventually revealed choice will falsify his prediction [because she actually chose to defect]. On the other hand, Adam has only slightly more favorable 53 to 47 odds of correctly forecasting Eve's cooperation when her eventually revealed choice will confirm his prediction [because she actually chose to cooperate].

Given the explicit calculation steps presented for the above examples, we omit further detailed calculations, and only summarize the relevant results similar to Tables 1 and 2 above, along with the associated strategy matrix. These results are shown in Table 3 and its associated strategy matrix.

**TABLE 3 & ITS STRATEGY MATRIX ABOUT HERE**

Note how the strategy matrix derived frm Table 3 implies ($CD_A$, $CD_E$) is a self-enforcing Nash equilibrium, despite players forecasting only slightly better than pure chance. This agrees with Theorem 1, since inequality (1) is satisfied; (.53)(25 - 4.8) - (.5)(25.2 - 5) = .606 > (5 - 4.8) = .2. Moreover, the ($CD_A$,$CD_E$) equilibrium is both "Pareto dominant" and "risk dominant"[6] over the ($DD_A$, $DD_E$) equilibrium. This means ($CD_A$, $CD_E$) satisfies Harsanyi & Selten's' "equilibrium selection" criteria, as well as many other such criteria. ($CD_A$, $CD_E$) may thus be the only really "stable" strategy equilibrium. Yet these conclusions follow even though players are able to forecast only slightly better than using a purely random device like flipping a coin to determine whether the other player will actually cooperate or defect.

In cases where both T - R and P - S are equal, it is easy to determine the threshold where examples like preceding one are possible. In particular, if the numerator of the cost-benefit ratio x/y satisfies, x = T - R = P - S, then inequality (1) is equivalent to:

$$r - w \geq \frac{x}{x + y} \qquad \text{where } y = R - P \tag{14}$$

In the preceding example, inequality (14) implies r - w must be at least (.2)/(15.2) ≅ .0099; which is satisfied for (r, w) = (.51, .5). Thus, only 51 to 50 odds of rightly rather than wrongly forecasting the other player's cooperation is sufficient for ($CD_A$, $CD_E$) to be a self-enforcing equilibrium in the preceding example.

6 Risk dominance follows because a 50-50 randomization between expected payoffs 15.3 and 4.9 exceeds a 50-50 randomization between expected payoffs 15.1 and 5; see pages 20-21 of, *Game Theory*, by Fudenberg & Tirole (1991). See also, *A General Theory of Equilibrium in Games*, by John Harsanyi & Reinhard Selten, MIT Press, (1988).

## V. INTUITIVE EXPLANATION FOR THE ($CD_A$, $CD_E$) EQUILIBRIUM

Recall implication (7b) that Eve's conditional choice probabilities and conditional forecast probabilities are necessarily equal whenever she conditionally cooperates in response to her forecast of Adam's actual choice [$CD_E$ implies ($u_E$, $v_E$) = ($r_E$, $w_E$)]. Consequently, if Eve can forecast Adam's actual choice better than pure chance [$r_E > w_E$], then $u_E > v_E$ is necessarily also implied by $CD_E$. That is, conditionally cooperating and positive forecasting ability together imply Eve is more likely to actually cooperate when Adam actually cooperates than when Adam actually defects. Moreover, the converse proposition also holds. That is, $u_E > v_E$ implies both $CD_E$ and $r_E > w_E$ must also hold [because responding according to $DD_E$ implies ($u_E$, $v_E$) = (0, 0) by implication (7a); and $r_E = w_E$ implies $u_E = v_E$ for either $DD_E$ or $CD_E$]. We thus have the following result.

THEOREM 3 (Characterizing When u > v Is Satisfied)

> u > v if and only if a player responds according to CD and can forecast the other player's actual choice better than pure chance (r > w). The absence of subscripts means this implication applies to both Adam or Eve.

COMMENTS:

1. We can use Theorem 3 to help explain why ($CD_A$, $CD_E$) can be a self-enforcing equilibrium. To do so, consider Figure 2 which shows a standard way of deriving the traditional dominant strategy PD-solution; corresponding to both players "always defecting" regardless of what they might forecast about each other's actual choice, ($DD_A$, $DD_E$). The two straight lines show Adam's expected payoff from actually cooperating (the lower line) or actually defecting (the upper line) as a function of Eve's unconditional probability of actually cooperating on the lower axis, $z_E = p(C_E)$. Since the two lines never cross, actually defecting is Adam's best response regardless of $z_E$.

**FIGURE 2 ABOUT HERE**

2. However, Theorem 3 implies there is not a *single* unconditional probability of Eve actually cooperating; if she can forecast Adam's actual choice better than pure chance and responds to her forecast according to $CD_E$. Rather, positive forecasting ability combined with conditionally cooperating (if and only if she forecasts Adam will cooperate) together imply Eve has *two* conditional choice probabilities, $u_E > v_E$. When this happens Adam's expected payoff from actually cooperating can be higher than from actually defecting, as illustrated by points E and F in Figure 2.

3. The reason for points like E and F is that $u_E > v_E$ implies Adam is no longer limited to exactly *vertical* comparisons of points on his two expected payoff lines (corresponding to a single unconditional probability $z_E$). Instead, Adam compares *non*-vertical points displaced by the horizontal difference between $u_E$ and $v_E$. Such comparisons are no longer guaranteed to favor actually defecting over actually cooperating. Similar reasoning applies to Eve when she compares her expected payoffs from cooperating versus defecting. She likewise is not limited to verticle comparisons if Adam can forecast her actual choice better than pure chance and responds to his forecasts according to $CD_A$; so that Theorem 3 also implies $u_A > v_A$.

4. Thus, if Adam and Eve can both forecast each other's actual choices better than pure chance, then conditionally cooperating may have higher expected payoff than always defecting for each of them, given each other also conditionally cooperates. This implication in turn implies that both players conditionally cooperating ($CD_A$, $CD_E$) may be a self-enforcing Nash equilibrium; as formally characterized in Part 2 of

Theorem 1 above. For example, if T - R equals P - S, then inequality (14) above implies ($CD_A$, $CD_E$) is a self-enforcing equilibrium whenever the difference between (u = r) and (v = w) exceeds x/(x+y) for both Adam and Eves' conditional choice and forecast probabilities. Otherwise, inequality (1) and its associated cooperation line can be used to determine which combinations of these probabilities imply ($CD_A$, $CD_E$) is a self-enforcing equilibrium.

## VI. FORECASTING ABILITY & ENDOGENOUS VS EXOGENOUS MESSAGE SOURCES

Let us further examine the statistical relationships between players' forecasts of each other's choices. Recall that a superscript f denotes a player's forecast of the other players' actual choice; so that $C_A^f$ and $C_E^f$ signify that Adam and Eve forecast each other will cooperate.

1. We can think of such forecasts as messages about what the other player's actual choice is expected to be. For example, $C_A^f$ might represent a message to Adam; such as "*Eve's actual choice is expected to be* $C_E$". Such a message may or may not be confirmed by Eve's eventually revealed choice. Consequently, it is up to Adam to decide whether to disregard this message (by always defecting regardless of Eve's expected choice, $DD_A$), or to respond to it in a contingent manner (by conditionally cooperating based on Eve's expected choice, $CD_A$). Similar strategic options ($DD_E$ or $CD_E$) are available to Eve in deciding how to respond to her expectations about Adam's actual choice.

2. Recall from equations (8c,d) that players' unconditional choice probabilities ($z_A$, $z_E$) are functions of their forecast probabilities and their response strategies, denoted $z^A[(r_A, w_A), (r_E, w_E); (XY_A, XY_E)]$ and $z^E[(r_A, w_A), (r_E, w_E); (XY_A, XY_E)]$. We can thus calculate players' unconditional probabilities of forecasting each others' cooperation (receiving messages $C_A^f$ and $C_E^f$ about each other) as follows:

$$p(C_A^f) = q^A[(r_A, w_A), (r_E, w_E); (XY_A, XY_E)] = r_A z_E + w_A(1 - z_E) \qquad (15a)$$

and

$$p(C_E^f) = q^E[(r_A, w_A), (r_E, w_E); (XY_A, XY_E)] = r_E z_A + w_E(1 - z_A) \qquad (15b)$$

$$z_A = z^A[(r_A, w_A), (r_E, w_E); (XY_A, XY_E)] = \frac{u_A v_E + v_A(1 - v_E)}{1 - (u_A - v_A)(u_E - v_E)} \qquad (15c)$$

where

$$z_E = z^E[(r_A, w_A), (r_E, w_E); (XY_A, XY_E)] = \frac{u_E v_A + v_E(1 - v_A)}{1 - (u_A - v_A)(u_E - v_E)} \qquad (15d)$$

and

$$\text{[DD implies (u, v) = (0, 0)] \& [CD implies (u, v) = (r, w)]; for either player} \qquad (15e)$$

3. We can use (15a,b,c,d,e) to calculate players' unconditional probabilities of them receiving messages ($C_A^f$,$C_E^f$), depending on their forecast probabilities and response strategies. For example, consider strategy pair ($CD_A$,$CD_E$); which implies by (15e) that (u, v) = (r, w) for both players. Substituting the latter equalities into (15c,d), and then into (15a,b), we obtain by algebraic manipulation the following formulas:

$$p(C_A^f) = \frac{r_A w_E + w_A(1 - w_E)}{1 - (r_A - w_A)(r_E - w_E)} \quad \& \quad p(C_E^f) = \frac{r_E w_A - w_E(1 - w_A)}{1 - (r_A - w_A)(r_E - w_E)} \qquad (16a,b)$$

Formulas (16a,b) are essentially the same as (15c,d), except that (r, w) is substituted for (u, v) for both players. This is because ($CD_A$,$CD_E$) implies both players cooperate exactly as often as they forecast each other's cooperation; so that their unconditional chance of cooperating equals their unconditional chance of forecasting each other's cooperation.

Next consider strategy pair ($DD_A$, $CD_E$). Using (15e) and substituting into (15c,d) implies ($z_A$, 1 - $z_A$) = (0,1) and ($z_E$, 1 - $z_E$) = ($w_E$, 1 - $w_E$); which in turn implies by (15a,b) that $p(C_A^f) = r_A w_E + w_A(1 - w_E)$ and $p(C_E^f) = w_E$. Similar reasoning applied to strategy pair ($CD_A$,$DD_E$) implies analogous results, except subscripts A and E are reversed; so that $p(C_A^f) = w_A$ and $p(C_E^f) = r_E w_A + w_E(1 - w_A)$.

Finally, consider strategy pair ($DD_A$,$DD_E$). Again using (15e) and substituting into (15c,d) implies ($z_A$, 1 - $z_A$) = (0,1) and ($z_E$, 1 - $z_E$) = (0, 1); which in turn implies by (15a,b) that $p(C_A^f) = w_A$ and $p(C_E^f) = w_E$. Table 4 displays the above calculations for different strategy pairs.

**TABLE 4 & TABLE 5 ABOUT HERE**

Notice from Table 4 that players' forecast-message probabilities [$p(C_A^f)$, $p(C_E^f)$] depend on their strategy combination. For example, the probability of Adam forecasting Eve's cooperation (by receiving message $C^f_A$) drops as either player shifts from conditionally cooperating to always defecting; dropping to [$r_A w_E + w_A(1 - w_E)$] if he unilaterally switches from $CD_A$ to $DD_A$, and dropping further to $w_A$ if Eve alone or both of them switch from $CD_E$ to $DD_E$. Table 5 gives two numerical examples where both players forecast probabilities are either (r, w) = (.9,.3), or (r, w) = (.51, .5). In the first case, message $C^f_A$ has a 75% chance of being received for strategy pair ($CD_A$, $CD_E$), dropping to a 48% chance for strategy pair ($DD_A$, $CD_E$), and dropping further to a 30% chance for strategy pairs ($CD_A$, $DD_E$) and ($DD_A$, $DD_E$).

5. Consider first the example with (r, w) = (.9, .3), and note the above mentioned drop from a 75% chance to a 48% chance of Adam receiving forecast-message $C_A^f$. Since he knows that Eve's forecast-message $C_E^f$ depends on the likelihood of his actual cooperation (because she is more likely to forecast his cooperation when he actually cooperates than when he does not), Adam knows Eve is less likely to cooperate if he decides to ignore his forecast-message and always defect instead. He is then less likely himself to forecast her cooperation (less likely to receive forecast-message $C_A^f$), because his likelihood of forecasting her cooperation also drops when doing so is less likely to be confirmed by her actual choice. Consequently, Adam anticipates that he will be less likely himself to forecast Eve's cooperation if he switches from conditionally cooperating to always defecting (from $CD_A$ to $DD_A$). The latter conclusion means that players realize the likelihood of their own forecast about each other depends on how they decide to respond to those forecasts. That is, the likelihood of receiving forecast-message $C_A^f$ depends on whether Adam ignores it or not.

6. Next consider the second example in Table 5 with (r, w) = (.51, .50). The differences between Adam's forecast-message probabilities for different strategy combinations are much smaller than in the first example. Nevertheless, as noted at the end of Section IV, these small differences are sufficient to enable just enough forecasting ability (above pure chance), so that ($CD_A$, $CD_E$) is still a self-enforcing Nash equilibrium when the cost-benefit ratio x/y = .01 [such as the case when (T, R, P, S) = (25.2, 25, 5, 4.8) in the earlier example]. Thus, what might seem like a negligible deviation from pure-chance forecasting can still make a major difference in the strategic opportunities available to players. For example, with the payoffs just mentioned, the two strategy equilibria ($CD_A$, $CD_E$) and ($DD_A$, $DD_E$) correspond to a difference of 15.11 versus 5 respectively in both players' expected payoffs.

7. The formulas in Table 4 and the numerical examples in Table 5 imply the likelihood of players receiving their forecasting-messages is ***not*** independent of how they respond to these messages. That is, $p(C_A^f)$ and $p(C_E^f)$ are not independent of players' strategy decisions [CD or DD]. Rather, the likelihood of a player's forecast-message itself "***endogenously***" depends on how *both* players' decide to respond to their own messages. Suppose players instead used an "*external*" or "***exogenous***" message source to forecast each other's cooperation. For example, suppose they used a naturally correlated message source such as temperature observations at different locations to forecast each other's cooperation (by each player picking a certain range of temperatures for which the other player's cooperation is forecasted if a temperature within this range is observed). Or suppose they set up an artificial message source by having an electronic random number generator send to each of them correlated signals; where players decide how they will respond to their own signal by either cooperating or defecting.

Consider a key feature of any such "exogenous" message source that might be used by Adam and Eve in a PD-situation. Let M denote the set of potentially observed signals from such a message source, with individual signals observed by Adam and Eve denoted $m_A$ and $m_E$. Suppose Adam decides to cooperate in response to certain signals, denoted $M_A \subset M$; and defect whenever a message $m_A \in (M - M_A)$ is received. Eve similarly selects a subset of potential messages $M_E \subset M$, and cooperates if and only if she observes a message $m_E \in M_E$. Subsets $M_A$ and $M_E$ thus correspond to Adam and Eves' forecasting messages $C^f_A$ and $C^f_E$; that is, $[C^f_A \Leftrightarrow m_A \in M_A]$ and $[C^f_E \Leftrightarrow m_E \in M_E]$. An exogenous information source implies *for any potential message received by Eve, $m_E \in M$, the probability of Adam receiving a message $m_A \in M_A$ is fixed independently of both Adam or Eves' response strategies, regardless of what actions such strategies might lead them to actually choose.* That is, $p(m_A \in M_A \mid m_E)$ is invariant to Adam or Eve's strategy decisions CD or DD, regardless of whether C or D is actually chosen by responding according to either strategy. For example, the correlation between thermometer readings at different locations is independent of what someone might do after looking at one of the thermometers, or what someone else might do after looking at the other thermometer.

8. The above property (of exogenous message sources) implies Eve can determine the likelihood of Adam observing $m_A \in M_A$ conditional on her observing $m_E \in M_E$. However, she cannot determine from $m_E \in M_E$ the likelihood of Adam actually cooperating or defecting in response to observing $m_A \in M_A$. That is, *the probability of Eve observing $m_E \in M_E$ conditional on Adam observing $m_A \in M_A$ is the **same** regardless of whether Adam actually cooperates or not in response to observing $m_A \in M_A$ [as well as the same regardless of how she actually responds to observing $m_E \in M_E$].* This is a general result for any information source whose signal-correlation is independent of players' response strategies which lead them to make actual choices; as stated in the next definition and theorem.

DEFINITION 16 (Forecast Probabilities For Exogenous Message Sources)

Let a set M represent any message source for which the conditional probability $p(m_A \mid m_E)$ between any pair of Adam and Eves' messages $m_A, m_E \in M$ is independent of both players' strategy decisions (CD or DD), no matter what actual choices (C or D) might actually result from these decisions. M is then said to be an *exogenous* message source; otherwise M is said to be an *endogenous* message source. Also, let $M_A$ and $M_E$ be any two subsets of M that Adam and Eve might use to forecast each other's cooperation; meaning $C_A^f \Leftrightarrow m_A \in M_A$, and $C_E^f \Leftrightarrow m_E \in M_E$. Their resulting conditional and unconditional forecast probabilities are denoted,

$$r_A(M_A) = p(m_A \in M_A \mid C_E) \qquad r_E(M_E) = p(m_E \in M_E \mid C_A)$$
$$\text{and} \tag{17a}$$
$$w_A(M_A) = p(m_A \in M_A \mid D_E) \qquad w_E(M_E) = p(m_E \in M_E \mid D_A)$$

$$z_A(M_A) = p(m_A \in M_A) \qquad \text{and} \qquad z_E(M_E) = p(m_E \in M_E) \qquad (17b)$$

THEOREM 4 (Exogenous Message Sources Imply Zero Forecasting Ability)

Let M be any exogenous message source. Then the resulting forecast probabilities defined in (17a) are equal to players' respective unconditional forecast probabilities defined in (17b); which thereby implies both players have zero ability to forecast each other's actual cooperation better than pure chance. That is, the following identities necessarily hold:

$$r_A(M_A) \equiv w_A(M_A) \equiv z_A(M_A) \quad \text{and} \quad r_E(M_E) \equiv w_E(M_E) \equiv z_E(M_E);$$

for any *exogenous* message source M, and for all $M_A$ Ì M, $M_E$ Ì M

COMMENTS:[7]

1. Theorem 4 implies any exogenous message source is a worthless guide to forecasting players' actual choices in a strategic PD-situation. This is an intuitive result, since rational players would not expect to forecast each other's strategic behavior successfully with messages that are only correlated with other messages, but not with their actual responses to any given message either of them might receive. For example, we might paraphrase the Nobel citation quoted at the beginning in the following way,

> "*Everybody knows in games like chess or poker, that players would not attempt to forecast expected countermoves from the other player by observing exogenous signals such as correlated temperature readings. Instead, they would try to discern signals correlated with each other's actual choices, or with those 'states-of-mind' which lead each of them to make actual choices*."

2. As suggested by the preceding statement, forecasting strategic behavior (better than pure chance) requires players to discern signals (endogenously) correlated with each other's actual "state of mind", or with whatever cognitive or behavioral mechanisms actually generate their behavior. For example, a human player's actual state of mind might be correlated with "facial expression," "body language," "tone of voice," and so on; or with its "perceived consequences" related to potentially received payoffs such as (T, R, P, S). Nonhuman players like birds and monkeys might be governed by relatively more "instinctive" mechanisms

7 Care must be given to avoid possible misinterpretation of Theorem 4. Suppose Adam forecasts Eve's cooperation if he observes temperatures above $90^0$ fahrenheit, and follows the response strategy of cooperating himself if and only if he observes temperatures satisfying this forecasting criteria. Given Adam's *assumed* response strategy, if he chooses to cooperate then he must have observed a temperature above $90^0$; which in turn implies Eve is likely to observe similar temperatures (because her temperature observations are correlated with his temperature observations). Thus, the likelihood of Eve's observed temperatures depends on Adam's actual choice, for any *given* response strategy that Adam might follow. On the other hand, the likelihood of Eve's observed temperatures is independent of whether Adam actually follows or deviates from any given response strategy. For example, the likelihood of it being hot (say above $90^0$) where Eve lives has nothing to do with how Adam might respond when it is also hot where he lives, or whether he might change the way he responds when he feels hot. Yet, forecasting Adam's actual choice requires Eve to forecast whether he will actually follow or deviate from any hypothetically given response strategy. If she cannot do the latter by "feeling hot", then such feelings are also worthless in forecasting Adam's actual choice; no matter how correlated their observed temperatures might be.

that also produce observable symptoms correlated with body language, facial expression; or with certain forms of "emotion", "vocal patterns", and so on.[8]

3. The above mentioned messages may depend on some type of "short-range" or "face to face" situation in order for such messages to be reliably perceived by the players involved. Even though they will be separated (without further communication) when they must actually choose between cooperating versus defecting, messages like those mentioned above may help them infer through "introspection" what each other is likely to do. Such "private" introspection may be mistaken, and would not necessarily have any "causal" influence on the other player's actual state of mind. Nevertheless, when combined with appropriate "short-range" prior communication, players' private introspections (while separated) might at least be more correlated than using a pure chance mechanism such as rolling a die; or trying to forecast each other's state of mind with exogenously correlated signals like temperatures observed at different locations.[9]

4. Another possibility might be endogenously correlated "focal points" originally discussed by Schelling (1960), and recently formalized by Sugden (1995). For example, players with with a common historical or cultural background might be able to use "most frequently mentioned" labels (1995, page 547) associated with cooperating or defecting to help correlate their individual expectations about when each other is likely to cooperate or not. Still another possibility might involve players exchanging "linguistic messages" about their attitude toward each other's cooperating or not. For example, Adam might say to Eve that only "scumbags" would cheat on somebody else[10]. Such linguistic messages may produce "emotional responses" within each player[11], which are correlated with those future states of mind (while players are separated) leading to their actual choices.

---

[8] Similar principles are used by "lie-detectors"; which measure small changes in physiological symptoms (like skin perspiration, pulse rate, or muscle tension) that are affected by a person's internal state of mind; such as an awareness of being truthful or not about one's past actions or future intentions.

9 Subjects in many PD-experiments report that introspection about themselves and the other subject's "point of view" played a role in their actual decisions. In doing so, subjects do not suggest their introspections have a "causal" effect on other subjects; just that they hope to get a handle on someone else's likely thinking through their own deliberations. Given the recurrence of such reports, it may be worth paying attention to them theoretically, rather than only analyzing the effects of using exogenous message sources (which cannot forecast better than pure chance, by Theorem 4).

10 Experiments of such examples are recently discussed by Elinor Ostrom (1994; "Frontiers of Research into the Design of Institutions," Seminar in Political Economy, John F. Kennedy School of Government, Harvard University, April 1994. Related research involving comparative field studies of American Indian cultures is also discussed by Stephen Cornell & Joseph P. Kalt (1995 a,b,c).

11 Emotional responses may also be correlated with changes in players' subjective evaluations of their basic payoffs. However, Theorems 1-3 above imply that potential changes in (T, R, P, S) are ***not*** necessary to affect a rational player's strategic incentives; because these theorems assume given payoffs unaffected by changes in players' ability to forecast each other's actual choices. Nevertheless, endogenous messages (especially those embodying a linguistic structure evolved from previous historical or cultural experience) may affect players' subjective beliefs about their individual payoffs, as well as their beliefs about the likelihood of each other actually cooperating; thereby leading to additional fruitful analysis about the resulting interdependence between these two effects. For example, as discussed in Section IX below, conditionally cooperative equilibria are sensitive to changes in the size of T - R, P - S, R - P ***even without reversing the ordinal PD payoff ranking T > R > P > S***. Consequently, a noticeable behavioral sensitivity in PD-situations may result from linguistic communication conveying ideas about "morality", "trustworthiness", social "roles" or "norms", etc; because such communication may alter the relative size of players' subjectively perceived payoff differences. If so, their equilibrium probability of cooperating can be affected ***without*** reversing the ordinal payoff ranking *T > R > P > S; as shown earlier by Theorem 2 and Corollary 1 above.*

5. Despite the above possible examples (of endogenously correlated signals), we do not wish to endorse any particular interpretation or theory about how players might forecast each other's strategic behavior (better that pure chance); except to say that something beyond exogenously correlated messages must be involved. Instead, we wish to investigate the theoretical implications of players of players having more than zero forecasting ability, however such ability might be achieved in practical situations with human or animal players. Theorems 1 and 2 above imply there is a "rationally self-enforcing incentive" for players to find some means of forecasting strategic behavior better than pure chance, even for one-shot PD-situations with the ordinal payoff ranking (T > R > P > S) still satisfied. Consequently, both humans and biological evolution may have found ways of perhaps systematically responding to this incentive, to the mutual advantage of the various kinds of players involved.

6. The sequel to this paper[12] links the above discussion (comments 1-4) to a large literature in experimental psychology directly concerned with the imperfect "detection" of signals, called *signal detection theory*.[13] Doing so enables one to construct a more general theory in which players' forecast probabilities ($r_A$, $w_A$; $r_E$, $w_E$) are not held fixed (as assumed in this paper). Instead, these probabilities are themselves "endogenously chosen" by the players in order to more effectively make use of (endogenous) messages that might help them improve their strategic forecasting ability.

## VII. COMPARISON WITH STANDARD THEORY & "CORRELATED EQUILIBRIUM"

Standard decision theory analyzes two general ways that players' actual choices might be statistically related to each other: "mixed strategies", and "correlated equilibrium" theory introduced by Robert Aumann[14]; as discussed in the following comments.

1. First consider mixed strategies. This means each player uses some kind of random process (such as flipping a coin or rolling a die), which is activated independently of the random devices used by other players. The signals resulting from using such devices are thus statistically independent of each other, with zero correlation between any pair of signals observed by different players independent of how they might actually respond to such signals. Thus, mixed strategies are a special case of players using an exogenous message source, whose signals are also totally uncorrelated with each other. Theorem 4 thereby implies any such message source is a worthless guide to forecasting players' actual strategic choices (better than pure chance). Hence, Corollary 1 above implies that players are involved in a "pure" PD-situation; where both of them always defecting on each other ($DD_A$, $DD_E$) is the only self-enforcing strategy combination for any ordinal payoff ranking, T > R > P > S. Moreover, Theorem 2 above implies that such behavior is completely invariant to changes in the cost-benefit ratio x/y associated with potential cooperation, so long as x/y > 0 [corresponding to T > R > P > S]. The latter theoretical prediction has been routinely contradicted in past experiments. Such experimental questions are further discussed below in Section IX, and in a sequel to this paper (Part III).

---

12 Part II, briefly outlined in Section X below.

13 See David Green & John Swets, *Signal Detection Theory and Psychophysics*, Robert Kreiger, New York, 1974; James Egan, *Signal Detection Theory and ROC Analysis*, Academic Press, New York, 1975; and John Swets, "Measuring the Accuracy of Diagnostic Systems," *Science*, **240**, June 3, 1988, pp. 1285-1293.

14 See Robert Aumann, "Subjectivity and Correlation in Randomized Strategies,"*Journal of Mathematical Economics*, 1974, **1**, pp. 67-96; and "Correlated Equilibrium as an Expression of Bayesian Rationality," *Econometrica*, January 1987, **55**(1), pp. 1-18. See also pages 53-60 of *Game Theory*, by Fudenberg & Tirole, MIT Press, 1991.

2. Next consider "correlated equilibrium" analysis. Here players have the opportunity to observe messages that may be highly correlated with each other, instead of being limited to uncorrelated signals as assumed for mixed strategies. However, any potential correlation between players' signals is still assumed to be invariant to their response strategies regardless of what choices may actually result from their strategies. Consequently, players have access only to exogenous message sources; which (by Theorem 4) still have zero ability to forecast their actual strategic choices. Thus, despite the correlation between different signals, such messages will not help players improve their forecasting ability above pure chance. Rationally self-interested players would therefore still always defect regardless of the relative costs versus benefits of potentially cooperating [so long as the cost-benefit ratio $x/y > 0$, by Theorem 2 and Corollary 1].

3. It is easy to verify that standard analysis, despite allowing correlated messages, nevertheless still assumes only exogenous message sources. The reader is invited to look at standard textbooks in game theory, and find those sections discussing correlated messages for one-shot simultaneous games. For example, see pages 316-319 of Binmore's (1992) textbook, *Fun and Games*. Often a text's formal definitions are illustrated by showing a matrix that contains joint probabilities of different combinations of players' forecast-message sets ($M_A$ and $M_E$), denoted $p(M_A, M_E)$, $p(M - M_A, M_E)$, $p(M_A, M - M_E)$, $p(M - M_A, M - M_E)$. For example, see Figure 7.17(b) on page 316 of Binmore's text. The message probabilities in this figure are assumed fixed, and used to derive probabilities of one player's messages conditional on the other player's potentially observed messages [by dividing individual message probabilities within each row or column of the matrix by the sum of the probabilities over that row or column respectively].

Deriving conditional message probabilities in the above manner guarantees they must all be independent of players' actual choices and their response strategies which produce such choices. Consequently, the implications discussed in Comments 1 and 2 above still hold, irrespective of the message correlation shown (but held exogenously fixed) in the message probability matrix. The same conclusion also applies whenever the formal equations used to define players' expected payoff calculations assume fixed message probabilities as players condition such calculations on their hypothetically making different choices.

4. For example, Aumann's formal definition and theorem characterizing a correlated-equilibrium (1987, pages 3-7)[15] assume fixed message probabilities irrespective of whether players actually follow or deviate from any given decision rule[16] for responding to their messages. He also discusses an abstract interpretation involving "Bayesian rationality" and "states of the world" defined so as to include each player's actual choice. No matter what interpretation might be suggested, Theorem 4 above about zero forecasting ability still applies so long as fixed message probabilities are assumed in the formal equations representing players' expected payoff calculations (corresponding to them hypothetically changing their response strategies). Consequently, an exogenous message source with zero forecasting ability is still

---

15 In particular, exogenously fixed message probabilities are assumed in Aumann's definition 2.1 on page 4 (defining a correlated equilibrium), and in proposition 2.3 on page 6 (characterizing which choice probability distributions are correlated equilibria); as well as in his "main theorem" on page 7 (linking Bayesian rationality to correlated equilibrium). Exogenously fixed message probabilities are also assumed in the figures Aumann uses to represent numerical examples of correlated equilibria; such as figures 2-5 on pages 4-5, and figures 7-8 on pages 14-16.

16 Notice as shown in Table 4 that players' message probabilities are not assumed given, and then used to determine their best choices. Instead, players' message probabilities are themselves derived as a partial consequence of their self-enforcing equilibria; which in turn endogenously depend on players' forecasting abilities (represented by their conditional forecast probabilities). In short, players' message probabilities are derived partly from their strategic decisions, instead of being assumed independently given and used to derive such decisions.

implicitly assumed in Aumann's Bayesian rationality interpretation of correlated equilibrium. The latter (Bayesian rationality) interpretation is further discussed in the Appendix.[17]

## VIII. FURTHER INTERPRETATION AND RELATED THEORY

### 1. Statistically Independent Choice Is A Degenerate Subset of Possible Choice Distributions

Recall what "*everyone knows*" according to the Nobel citation quoted at the beginning; namely, "*players have to think ahead and devise a strategy based on the expected countermoves from the other player.*" As also discussed above, this strategic perspective implies that players have a self-enforcing incentive to forecast each other's "expected countermoves" better than pure chance; that is, in forming such expectations when they are more likely to be confirmed rather than falsified by each other's eventually revealed choices. To the extent they achieve this objective , Theorem 1 implies ($CD_A$, $CD_E$) may be a self-enforcing equilibrium for certain one-shot PD-situations; which in turn implies by Theorem 3 that their actual choices may thereby ***not*** be statistically independent.

Consequently, the strategic interdependence embodied in one-shot PD-situations may motivate rationally self-interested players to forecast each others' behavior well enough to imply their actual choices are not statistically independent. Actually succeeding in doing so may not be easy, but there is a "natural" incentive in trying to do so, because the ($CD_A$, $CD_E$) equilibrium which thereby might be achieved Pareto dominates the ($DD_A$, $DD_E$) equilibrium.

More generally, we suggest there is likewise a natural intuition for rationally self-interested players believing that *strategically* interdependent choices may also lead to *statistically* interdependent choices; not just for one-shot PD-situations, but potentially for any situation where strategically interdependent consequences exist. For example, as modeled in Theorems 1-3 above, strategically interdependent consequences may imply (self-enforcing) "forecasting-incentives" motivating rational players to make statistically interdependent choices. Given such possibilities, it may not be fruitful to impose an a priori restriction against theoretically investigating statistically interdependent behavior. Yet, many presentations of noncooperative game theory allow only statistically independent mixed-strategies into the formal analysis.[18]

To get a further perspective of what is involved in such a methodological restriction, consider the joint probability distribution of players' actual choice combinations, denoted [$p(C_A, C_E)$, $p(D_A, C_E)$, $p(C_A, D_E)$, $p(D_A, D_E)$]. Let S denote the set of all joint-probability distributions [all those distributions whose individual probabilities are non-negative and sum to one]. Note that these joint-probabilities are required to determine the likelihood of any specific payoff (T, R, P, S) actually being received by a player. This is because both players' actual choices are necessary to determine the specific payoff received by either player. Consequently, only the joint probability distribution of players' actual choices ultimately matters to Adam

---

17 Part 2 of the Appendix shows that there is an inconsistency in Aumann's Bayesian interpretation; which when resolved implies that Bayesian rationality is consistent with players using either exogenous or endogenous information sources.

18 For example, Binmore's 1992 game theory textbook (*Fun and Games*) states on pages 232-233: "A standard assumption for a noncooperative analysis ... is that the random devices used by the players in implementing their mixed strategies are *independent*. This should always be assumed to be the case unless something is said to the contrary."

or Eve when they evaluate the desirability of alternative strategies like CD versus DD (conditionally cooperating versus always defecting).

Now consider the following relationship discussed by Fishburn (1982, pages 87-88): any joint-probability distribution generated by statistically independent choices must satisfy,[19]

$$p(C_A, C_E)p(D_A, D_E) = p(C_A, D_E)p(D_A, C_E) \qquad (18)$$

Only a subset of "measure zero" within S can satisfy equation (18); which means "almost all" joint-probability distributions *cannot* be generated from statistically independent behavior. Consequently, statistically independent behavior represents a "degenerate subset" within S. This implication suggests the narrow range of potential cases one is limited to by not investigating situations where strategic interdependence also gives rise to statistically interdependent behavior.

## 2. Statistically Interdependent Choice and the Axioms of Expected Utility

The above discussion also relates to using expected utility maximization to represent the meaning of rational behavior under uncertainty. In particular, the original expected utility axioms were applied to nonstrategic situations; where events beyond a decision maker's control affected its achieved utility. For example, uncertain weather conditions may affect the future yield from planting a particular crop. A plausible assumption for such situations is that nonstrategic events are statistically independent of actions taken by the decision maker. Such independence has been a key ingredient in formal axiomatizations of expected utility since the early proofs of Savage and von Neuman-Morgenstern.

However, when applied to game theory situations, other "events" may result from the "expected countermoves" of other players. Such intrinsically strategic events may no longer be statistically independent of any particular player's decisions, as suggested above. Nevertheless, a major reason for assuming statistically independent behavior is to preserve the formal justification for maximizing expected utility in explicit game theory settings. How then is one to allow for the possibility of statistically interdependent choices without giving up an axiomatic justification of expected utility theory in the process?

The following answer is suggested; namely, *to assume players each have a subjective preference ordering applied directly to the set of potential joint-probability distributions* S. The reason for doing so is that, as noted in the preceding section, only joint probabilities of players' actual choice combinations ultimately determine the likelihood of Adam or Eve receiving any specific payoff, and hence the desirability to them of any strategically generated uncertainty about which specific payoffs they will ultimately receive.

Expected utility axioms can then be applied directly to players' subjective preferences over S; thereby motivating players to maximize the expected utility of those joint probability distributions which are feasible under different strategic situations (rather than limiting theoretical investigation to a degenerate subset within S). In this way we can formally justify the use of expected utility theory[20] without presupposing the nature of statistical interdependence resulting from different strategic situations. Instead, such situations can be left open to determine whatever behavioral and statistical relationships are consistent with the incentives, decision-making abilities, and forecasting abilities of the players involved.

---

19 Statistically independent choices imply each joint-probability equals the multiple of the corresponding unconditional choice probabilities: $p(C_A, C_E) = z_A z_E$, $p(C_A, D_E) = z_A(1 - z_E)$, $p(D_A, C_E) = (1 - z_A)z_E$, $p(D_A, D_E) = (1 - z_A)(1 - z_E)$. These equalities immediately imply equality (18).

20 The main theorems of this paper also hold if "non-expected utility theories" are used instead of traditional expected utility; such as Fishburn's SSB theory; see Machina (1985).

3. Revealed Preference Theory & Cooperation versus Defection In PD-Situations

The traditional intuition about defecting always being the best strategy in PD-situations (discussed in the introduction) has been so strong that some theorists have argued this intuition is a logical tautology implied from the very meaning of players' "revealed preferences". For example, Binmore has recently endorsed this conclusion (1994, pages 104 - 107). Given that conditionally cooperating can be a self-enforcing Nash equilibrium by Theorem 1 above, we briefly explain why always defecting is not a tautological result of revealed preference theory. Consider in particular Binmore's argument, which combines revealed preference theory with a version of Savage's "sure-thing" principle.

1. Following Binmore (page 105), we write two revealed preference statements about how Adam will actually choose between two options s and t depending on his knowledge about the truth or falsity of some proposition P.

R1 Adam chooses s over t when he knows P remains true no matter what he chooses.

R2 Adam chooses s over t when he knows P remains false no matter what he chooses.

Thus, Adam will reveal (by actually choosing) a preference for s over t when he knows the validity of P (either true or false) is independent of what he chooses. But what if the validity of P itself depends on what Adam chooses; not necessarily totally, but perhaps partially in that the likelihood of P being true depends on Adam's revealed choice? For example, suppose P is more likely to be true if Adam chooses s rather than t. Consider then a third revealed preference statement.

R3 Adam chooses s over t when he knows the likelihood of P being true depends of what he chooses.

Notice that statement R3 allows the validity of P to depend on what Adam chooses, but statements R1 and R2 do not. Consequently, R3 cannot be a tautological implication of R1 and R2.

2. Now apply the last conclusion to PD-situations by letting options s and t refer to Adam's choosing to defect and cooperate respectively, and letting P denote the proposition, "Eve chooses to cooperate." The reason for doing so is that if Eve conditionally cooperates with positive forecasting ability, then she will cooperate with greater probability when Adam actually chooses cooperation over defection (because $CD_E$ and $r_E > w_E$ imply $u_E > v_E$, by Theorem 3 above). Thus, with positive forecasting ability, the validity of P (meaning the likelihood of Eve choosing to cooperate) may in fact depend on what Adam actually chooses. Consequently, statement R3 applied to Adam choosing in a PD-situation cannot be a tautological result of statements R1 and R2 (which assume Eve's choice to either cooperate or not is the same no matter what Adam actually chooses).

3. The above comments suggest a way of rewriting Binmore's statement of the sure-thing principle (page 107) so as to be consistent with revealed preference theory, as shown next. The phrases in the square brackets are added to his statement in order to make explicit what is required for it to be consistent with revealed preference statements R1 and R2.

> "If Adam will choose to defect both when he knows Eve will choose to cooperate [no matter what he chooses], and when he knows Eve will not choose to cooperate [no matter what he chooses]; then the sure-thing principle says he must also choose to defect no matter what he might believe about the likelihood of her prospective choice [provided he also believes her likelihood, whatever it might be, is the same no matter what he chooses].

The above statement makes clear that the sure-thing principle does not imply Adam will always defect in PD-situations, except at the limit of zero forecasting ability. This is because ($r_E = w_E$) is the only case where Eve's likelihood of cooperating is independent of Adam's choice even if she conditionally responds to her forecast (that is, even if she responds according to $CD_E$). Consequently, it is not a tautology (either from revealed preference theory or from the sure-thing principle) that rational players will necessarily choose defection in PD-situations, except at the limit where they cannot forecast each other's actions better than pure chance.

## IX. PREDICTING BEHAVIORAL SENSITIVITY TO PD-PAYOFFS: PAST EXPERIMENTS

1. Recall the intercept formulas for the cooperation line in Figure 2, which are shown below for convenience. These intercept values for r and w are denoted with a superscript *, and are written without subscripts A or E because they are the same for both players if their basic payoffs (T, R, P, S) are the same.

$$r^* = \frac{R - P}{(R - P) + (T - R)} \qquad \& \qquad w^* = \frac{P - S}{(P - S) + (T - R)} \tag{19}$$

As already discussed, these formulas imply the cooperation line will shift arbitrarily close to the $45^0$-line (in the unit probability box) corresponding to zero forecasting ability, as the two payoff differences T - R and P - S get sufficiently small relative to R - P. This possibility also corresponds to the cost-benefit ratio for cooperating in PD-situations x/y dropping sufficiently close to zero. Consequently, conditionally cooperative equilibria ($CD_A$, $CD_E$) necessarily come into play as x/y drops sufficiently, so long as players can forecast each other's actual choices better than pure chance.

2. The last conclusion implies that for any given forecasting skill better than pure chance (represented by $r_A > w_A$ and $r_E > w_E$), cooperative choices resulting from ($CD_A$, $CD_E$) are more [less] likely to happen as the cost-benefit ratio x/y gets smaller [larger]. Thus, Theorem 1 implies that conditionally cooperative behavior will be sensitive to changes in players' basic payoffs even when the ordinal ranking T > R > P > S is preserved. Moreover, such "behavioral sensitivity" is directly related to the two payoff differences T - R and P - S compared to R - P. This general pattern (*which is now a formally derived consequence of rationally self-interested behavior*) agrees with the results of many earlier experiments.[21]

In fact, this pattern has been so widely observed that researchers (in economics, political science, sociology, psychology, etc.) have coined terms to suggest why behavioral sensitivity to these payoff differences is "intuitively likely" to happen. For example, the term "greed" is associated with the potential "net profit" T - R from taking advantage of someone's cooperation; and the term "fear" is associated with the "aversion" to potentially loosing P - S compared to a guaranteed minimum payoff P from defecting. A recent survey of the literature concerning behavior in PD-situations by Liebrand et, al (1992, pages 15-17) suggests the systematic nature of this observed pattern.

> "There is strong support for the generality of influence of different payoff structures. ... payoff structure has been shown to be strong not only in experimental work, but also in field studies. ...two motives, namely fear and greed, may lead one to choose noncooperatively ... both motives are important. People are more likely to cooperate to the extent that they loose less by cooperation when others do not cooperate. Similarly, people are more likely to cooperate to the extent that they gain less by taking advantage of, or free riding on, the cooperation of others."

---

21 See for example, Liebrand et. al. (1992), Rapoport (1966), Coombs (1973), Strobe & Frey (1982), Simmons, Dawes, & Orbell (1984)

3. Despite the above "generally observed" patterns, traditional analysis of PD-situations has tended to either ignore them (by arguing the experiments do not adequately simulate "real" PD-situations) or categorize these patterns as illustrating "irrational" behavior (because zero behavioral sensitivity is the only rational solution permitted by traditional analysis). Theorems 1,2 and Corollary 1 above enable a different interpretation which predicts behavioral sensitivity to "fear" and "greed" motives as a consequence of rational self-interest.

4. Another hypothesis relates to these theorems. Namely, preplay communication may help players improve their ability to forecast each other's behavior after they are separated, but before choosing themselves to cooperate or defect while separated (thereby raising r relative to w for each player).[22] If so, then conditionally cooperative equilibria are more likely to arise for any given payoff structure [because players' (r, w) probabilities are more likely to be on or above any given cooperation line implied by their payoff structure]. We would then see a positive link whereby more "preplay" communication stimulates more "within play" cooperation (even without any binding commitments being enforceable once players are separated). This general pattern also agrees with many experiments.[23] Much of this data can thus also be explained as the predicted consequence of rational self-interest.

5. A number of other predictions can be calculated and implemented experimentally with the intercept formulas (19). This is because changes in the payoff structure may not only shift the cooperation line in or out, but also simultaneously rotate it in either direction. We can thereby use equations (19) to calculate predictions that have not been tested before. These are briefly discussed in the next section. For now the main point is that we are no longer limited to theorizing in a manner that cannot explain any behavioral sensitivity in one-shot PD-situations, so long as the ordinal payoff ranking (T > R > P> S) is preserved.

## X. FURTHER THEORY AND EXPERIMENTS: PARTS II & III

Preceding analysis focused on symmetric PD-games in order to show why conditional cooperation can be a self-enforcing Nash equilibrium in a relatively simple but rigorous manner [despite the "dominant strategy" intuition discussed in the introduction, or the "revealed preference" intuition discussed in Section VIII(3)]. Subsequent papers develop the analysis without players' payoffs or forecast probabilities being the same; showing that similar results hold in the general case. These papers are briefly outlined next.

Part II. Nash Equilibria With Endogenous, Jointly-Feasible Forecast Probabilities

1. This paper allows players' forecast probabilities to differ, and be "endogenously" varied by each player. The analysis draws on a large literature in behavioral psychology about imperfect detection of messages, called "signal detection theory".[24] Doing so ensures that assumptions about the statistical nature of players' forecasting skills are not arbitrarily specified, but instead conform to empirically verified regularities extensively studied by experimental psychologists.

---

22 This may happen because preplay communication may involve messages endogenously correlated with players' states of mind when they actually choose (rather than exogenous messages), as discussed earlier in Section IV. A more "intimate", "face to face" preplay setting may help players identify or "detect" such endogenous messages.

23 See for example, Orbell, van de Kragt, & Dawes (1988), Nisbet & Wilson (1977), Radnoff & Weidner (1966), Dawes, McTavish, & Shakley (1977).

24 See the references in footnote 12 above; as well as Davies (1969), Egan (1967), McNicol (1972), Schulman & Greenberg (1970), Swets & Pickett (1982), and Tanner, Rauk, & Atkinson (1970).

2. Because of the strategic interdependencies involved in simultanous games, one player's forecast probabilities cannot be arbitrarily selected independent of another player's forecast probabilities. Nevertheless, there is a well defined structure of "jointly-feasible" forecast probabilities in one-shot PD-situations which can be precisely characterized. Doing so shows that players can independently select their forecasting strategies provided they are jointly feasible, and that similar results to Theorem 1 necessarily hold without any further restrictions.[25]

3. Other theoretical topics are also discussed. These include an analysis of certain questions such as "Newcomb's paradox"; showing that better forecasting ability brings with it an unavoidable instability in the probability of cooperating near the limit of perfect forecasting. As a result, stable cooperation may not result from ever more accurately correlated choices in a PD-situation. Similar analysis also applies to the "twin paradox".

4. One can also incorporate the opportunity cost of players choosing to improve their strategic forecasting abilities; showing for example that it is always worth forecasting well enough to sustain conditionally cooperative equilibria as the cost-benefit ratio x/y gets small enough. On the other hand, weak assumptions imply that it is never worth forecasting well enough to sustain conditionally cooperative equilibria as the cost-benefit ratio x/y gets sufficiently large.

Part III. Empirical Implications and New Experiments

As suggested in the above title, this paper focuses on empirical and experimental issues. If players' individual payoffs are not equal, then two cooperation lines (one for each player) result from the analysis. These can be independently varied relative to each other, resulting in new experimental cases. Other experiments involve "boundary cases" near the limit where the ordinal payoff ranking ($T > R > P > S$) is reversed. Such experiments help one determine the relative predictive power of alternative theories of cooperative behavior in PD-situations. None of these predictions would be theoretically possible (that is, for rationally self-interested players) without investigating the strategic implications of positive forecasting ability, and endogenous versus exogenous message sources.

## XI. CONCLUSION: THE HISTORICAL IMPORTANCE OF ONE-SHOT COOPERATION

Traditional analysis shows that cooperation can be much easier to achieve once a one-shot PD-situation is repeated; especially when such repetition might continue indefinitely. Perhaps most currently observed cooperation in the field is due at least in part to the incentive effects of long term, "recurrent" relationships. Even if this is actually the case, one-shot relationships may still have an essential role to play in the development of cooperative individual behavior and social institutions. Consider what might happen with no possibility of successful one-shot cooperation, and ask the following question. Namely, would repeated relationships get started and continue robustly without the possibility of successful one-shot relationships to initiate them, or to sustain them if something happens that temporarily interrupts them? Even if we could imagine players somehow "skipping" directly to stable long-run relationships without any successful one-shot relationships, what would happen if the players involved lacked the cognitive ability to conceive of consequences beyond the near future?

---

25 The basic result is that conditionally cooperative strategy combinations ($CD_A$, $CD_E$) can be self-enforcing Nash equilibria if and only if both players' jointly feasible forecast probabilities can lie on or above the cooperation line; so that inequality (1) is satisfied.

Think of these questions especially in terms of an historical progression involving numerous self-interested individuals who interact over a long succession of relatively short-run situations. Recall the earlier discussion about endogenous (rather than exogenous) messages being necessary to forecast another player's actual choices better than pure chance. Recall also that such messages may be discernable through various forms of "close range" or "face to face" communication involving "body language", "facial expression", "emotion", "vocal patterns", and so on. Such close range communication has been systematically evidenced in a number of nonhuman species, especially primates.[26] Most biologists also believe that animal communication is selfishly motivated.[27] Moreover, such communication also routinely arises either in actual short run situations, or in situations where the individuals involved lack sufficient cognitive ability to "self-recognize" more than the near future.

Consequently, a selfishly rational potential incentive toward cooperation in these situations may become historically significant when augmented with enough short-range communication ability to overcome the opposite incentive (by improving individuals' skill at detecting signals "endogenously correlated" with their actual choices; so that conditionally cooperative behavior becomes a self-enforcing Nash equilibrium). Theorems 1 and 4 may thus explain one of "nature's secrets"; namely, a way of motivating rationally self-interested cooperation even in the most difficult one-shot type PD-situations that typically arise ***before*** extended social cooperation evolves. Without this potential incentive (characterized by Theorem 1), and the communication skills needed to utilize endogenous messages (required by Theorem 4 to achieve positive forecasting ability), the kind of longer-run exchange and organization relationships[28] eventually typical of human societies may never have evolved in the first place.

---

[26] See for example, Bonner (1980), Smith (1977), and Wilson (1975).

[27] See for example, *The Selfish Gene*, Richard Dawkins (1976).

28 Including both market trading and contract relationships (hierarchy, hybrid, network) within nonmarket organizations studied in "transaction cost" analysis (following Coase, Alchain, Williamson).

**APPENDIX**

All of the theorems besides Theorem 1 either are proven in the main text, or follow easily from the definitions which apply to them. Thus, only Theorem 1 is proven next.

Part 1. Proof of Theorem 1

Recall we are dealing with symmetric PD-situations where both player's conditional forecast probabilities are equal [ so that $(r_A, w_A) = (r_E, w_E) = (r, w)$]; and start with determining the four expected payoff levels for Adam depending on whether and Eve conditionally cooperates and/or always defects. If they both always defect, then both are guaranteed to get payoff P; so we thus have,

$$U_A = p^A[DD_A, DD_E, (r, w), x = (T, R, P, S)] = P \qquad \text{(A1a)}$$

Next assume Adam conditionally cooperates while Eve always defects [strategy pair $(CD_A, DD_E)$]. Using implication (7b) and (7a) for Adam and Eve respectively implies, $(u_A,v_A) = (r, w)$ and $(u_E, v_E) = (0, 0)$. Next substitute these into equations (5a,b) to obtain players' unconditional probabilities of cooperating, $(z_A, z_E) = (w,0)$; and then substitute these values along with $(u_A,v_A) = (r, w)$ and $(u_E, v_E) = (0, 0)$ into the left-hand probability multiples of equations (3a,b,c,d) to obtain players' distribution of choice probabilties, $[p(C_A, C_E), p(D_A, C_E), p(C_A, D_E), p(D_A, D_E)] = [0, 0, w, 1 - w]$. Then substitute these values into equation (2a) to obtain Adam's expected payoff formula,

$$U_A = p^A[CD_A, DD_E, (r, w), x = (T, R, P, S)] = wS + (1 - w)P \qquad \text{(A1b)}$$

A similar sequence of substitutions [using (7a) or (7b) depending on which player always defects versus conditionally cooperates respectively] obtains the following formulas, for Adam's expected payoffs for strategy pairs $(DD_A, CD_E)$ and $(CD_A, CD_E)$:

$$p^A[DD_A, CD_E, (r, w), x = (T, R, P, S)] = wT + (1 - w)P \qquad \text{(A1c)}$$

and

$$p^A[CD_A, CD_E, (r, w), x = (T, R, P, S)] = zrR + z(1 - r)S + (1 - z)wT + (1 - z)(1 - w)P$$

$$= z[rR + (1 - r)S] + (1 - z)[wT + (1 - w)P] \qquad \text{(A1d)}$$

where $$z = \frac{rw + w(1 - w)}{1 - (r - w)(r - w)} = \frac{w}{1 - (r - w)} = \frac{w}{1 - r + w} \qquad \text{(A1e)}$$

Note that similar reasoning implies the same formulas for Eve's expected payoffs, by reversing the A and E subscripts in (A1a,b,c,d), and using the right-hand probability multiples of equations (3a,b,c,d).

With these formulas, the proof of Theorem 1 follows directly from a simple property of convex combinations applied to two amounts X and Y; namely,

$$X \geq Y \text{ if and only if } p(X)X + [1 - p(X)]Y \geq Y; \quad \text{for any } p(X) \geq 0 \qquad \text{(A2a)}$$

Note that (A2a) is also equivalent to,

$$Y > X \text{ if and only if } Y > p(X)X + [1 - p(X)]Y; \quad \text{for any } p(X) > 0 \qquad \text{(A2b)}$$

From this point on, we assume $w > 0$, in order prove the traditional result that DD *strictly* dominates CD (for both players), if inequality (1) is violated. Otherwise, DD only weakly dominates CD if $w = 0$. The latter case,

w = 0, does not affect Part 2 of Theorem 1, about CD also being a Nash equilibrium when inequality (1) is statisfied [because (1) requires only the weak inequality ³ instead of a strict inequality >]. The traditional dominant strategy result is equivalent to the above formulas satisfying, (A1a) > (A1b) and (A1c) > (A1d) for Adam's expected payoffs, and similarly for Eve's expected payoffs [by switching the A and E subscripts in (A1a,b,c,d)]. Thus, the traditional dominant strategy result holds if the following inequalities hold [also applying to Eve's expected payoffs by switching subscripts]:

$p^A[DD_A, DD_E, (r, w), x] > p^A[CD_A, DD_E, (r, w), x]$ Û $P > wS + (1 - w)P$, for $w > 0$ (A3a)

$p^A[DD_A, CD_E, (r, w), x] > p^A[CD_A, CD_E, (r, w), x]$ Û

$$wT + (1 - w)P > z[rR + (1 - r)S] + (1 - z)[wT + (1 - w)P] \quad \text{(A3b)}$$

for $$z = \frac{w}{1 - r + w} > 0 \quad \text{(A3c)}$$

Notice that inequality (A3a) follows from implication (A2b) by letting X = S, Y = P, and p(X) = w. Morever, inequality (A3b) also corresponds to an example of implication (A2b); where X = [rR + (1 - r)S], Y = [wT + (1 - w)P], and p(X) = z. Thus, inequality (A3b) is equivalent to determining whether the following inequality holds:

$$[wT + (1 - w)P] > [rR + (1 - r)S] \quad \text{(A4)}$$

Algebraic manipulation of inequality (A4) directly yields the following inequality, which is the *opposite* to the first version of inequality (1) in Definition 7 of the main test. That is, inequality (A4) is equivalent to P - S > r(R - S) - w(T - P); so that inequality (A4) is equivalent to (r, w) *violating* inequality (1). Thus, violating inequality (1) implies [along with inequality (A3a), which was also just shown to hold] that $DD_A$ is a dominant strategy for Adam. Similar reasoning also implies [by reversing subscripts A and E] that $DD_E$ is likewise a dominant strategy for Eve. Hence, (r, w) violating inequality (1) implies strategy pair ($DD_A$, $DD_E$) is the unique dominant strategy Nash equilibrium; which proves Part 1 of Theorem 1.

Next consider what happens when inequality (A4) is reversed, so that [rR + (1 - r)S] ³ [wT + (1 - w)P]; which is equivalent to (r, w) *satisfying* inequality (1). This is an example of implication (A3a) by letting X = [rR + (1 - r)S], Y = [wT + (1 - w)P], and p(X) = z; which in turn implies by formulas (A1c,d) that $p^A[CD_A,CD_E,(r, w),x]$ ³ $p^A[DD_A, CD_E, (r, w), x]$. Similar reasoning implies the same inequality also applies to Eve's expected payoffs [by reversing subscripts A and E]. That is, (r, w) satisfying inequality (1) also implies $p^E[CD_A, CD_E, (r, w), x]$ ³ $p^E[CD_A, DD_E, (r, w),x]$.

The latter two inequalities mean that satisfying inequality (1) implies that CD has higher expected payoff than DD (for either Adam or Eve) if the other player chooses DD. That is, (r, w) satisfying inequality (1) implies strategy CD is an optimal decision for either player from strategy set S = {CD, DD}, given the other player also selects strategy CD. Thus, (r, w) satisfying inequality (1) implies ($CD_A$, $CD_E$) is a Nash equilibrium.

Recall also that, as shown above, inequality (A3a) is guaranteed for any w > 0, and holds weakly if w³0 [regardless of whether inequality (1) holds]. A similar inequality to (A3a) also holds for Eve by reversing subscripts A and E [regardless of whether inequality (1) holds]. Thus, inequality (1) implies ($DD_A$., $DD_E$ ) is still a Nash equilibrium. Hence, (r, w) satisfying inequality (1) implies ($CD_A$,$CD_E$) and ($DD_A$., $DD_E$ ) are both Nash equilibria; which proves Part 2 of Theorem 1.

Part 2. Aumann's Bayesian Interpretation of Correlated Equilibrium

We now further discuss Aumann's interpretation that Bayesian rationality implies that the probability distribution of players' choice combinations is a correlated equilibrium. There is an inconsistency in his interpretation, which is discussed in points 8-9 below, and resolved in point 10 (showing that Bayesian rationality is consistent with players using either exogenous or endogenous information sources, as discussed in Section VI of the

main text). To do so, the following definitions are introduced for at least two players i = (1, ... n) deciding how to respond to information about other players' choices besides their own choice.

1. W denotes the set of all conceivable "states of the world"; with individual states denoted w Î W. $A_i$ denotes player i's action partition of W; meaning the set of all distinct actions for player i. Each element of $A_i$, denoted $a_i$ Î $A_i$, is a subset $a_i$ Ì W; corresponding to a smallest distinguishable action for player i. $M_i$ denotes player i's message partition of W; meaning the set of all distinct messages that player i can observe. Analogous to $A_i$, each element $m_i$ Î $M_i$, is a subset $m_i$ Ì W; corresponding to a smallest observable message for player i.

2. $A = A_1 \times ... \times A_n$, and $a = (a_1, ... , a_n)$ Î A; where a denotes an n-tuple of players' actions, called an action profile. Similarly define $M = M_1 \times ... \times M_n$, and $m = (m_1, ... , m_n)$ Î M; where m is an n-tuple on players' messages, called a message profile. Let G denote a countable probability space over W; where s Î G represents a probability distribution over all potential states in W. s is called a "state distribution".

3. Let P = M ´ A denote the set of all potential message and action profiles; with individual pairs of message/action profiles denoted p = (m, a) Î P = M ´ A. For any subset X Ì W representing an event [including player i choosing an action $a_i$ Î $A_i$ or observing a message $m_i$ Î $M_i$], the probability of X implied by s is denoted, $\sigma(X) = \sum_{\omega \in X} \sigma(\omega)$. For any collection of two or more sets [such as X, Y, Z, ...], let Ç(X, Y, Z, ...) = (X Ç Y Ç Z Ç ...) denote the intersection of these sets. Then the joint probability of a collection of events happening simultaneously is denoted,

$$\sigma(X, Y, Z, \dots) = \sum_{\omega \in \cap(X,Y,Z,\dots)} \sigma(\omega) \qquad \text{(A5)}$$

For example, s[p = (m, a)] is the probability of a pair of players' message/action profiles p = (m,a) Î P; meaning the probability of states w contained in the intersection of all the message and action events ($m_j$ Î $M_j$ and $a_j$ Î $A_j$] associated with a particular message/action profile p = (m, a). From a Bayesian perspective, s can be thought of as a common "prior" probability distribution of players' beliefs about the likelihood of potential states w Î W; including the likelihood of potential combinations of observation/action profiles simultaneously arising. If we wish these beliefs to be subjective to particular players, then each $s_i$ denotes player i's subjective prior distribution of potential states w Î W.

4. $(P, G, s_i)$ is called player i's "subjective information model"; which describes player i's subjective beliefs about the likelihood of all potential pairs of players' observation/action profiles pÎ P.

5. Following Aumann, consider an "outside observer" perspective, where each player i must evaluate how best to respond to its observed messages or "signals" $m_i$ Î $M_i$. Quoting Aumann (1987; page 8),

> In analyzing the situation, each player i ... cannot ignore the possibility of his receiving a signal different than the one he actually got, even though he knows that he did not actually get such a signal. This is because the *other* players do not know what signal he got. Player i must take the ignorance of the other players into account when deciding on his own course of action, and he cannot do this if he does not explicitly include in [his information] model signals other than the one he knows he got.

6. Also following Aumann (1987; page 7), consider what happens when each player knows nothing beyond observing himself choose an action; so that his action $A_i$ and message $M_i$ partitions of the state space W coincide. That is, $M_i = A_i$ for all i; meaning $m_i = a_i$ are indistinguishable events for each player i; which in turn means M = A. Then apply this knowledge situation to the above quotation; so that the "signals" referred to in the quotation represent each player i observing himself choosing action $a_i$ [because player i's actions "are" his signals; $a_i = m_i$ for all $m_i$ Î $M_i$]. In such a situation, player i knows that although the other players do not know what specific action he chose, *they nevertheless know he cannot have chosen differently than what he actually observed himself choose (whatever his observed action might have been*). Consequently, player i knows that all the other players assign zero probability to any such "conceivable event", corresponding to his choosing differently than he actually observed himself choosing.

7. The latter conclusion is formally implied for any subjective information model $(P,G,s_i)$ over a state space W. To see this, note that M = A implies P = M ´ A = A ´ A; so that there exist "conceivable" or "hypothetical"

observation/action profiles (a, a¢) Î P such that a ¹a¢ because $a_i$ ¹ $a_i$¢ for some player i. The latter inequality implies the intersection Ç(a, a¢) = Æ [because the elements of a partition $A_i$ are disjoint subsets whose intersection is therefore empty]; which in turn implies from equation (A5) that $s_i$(a, a¢) = 0 whenever a ¹ a¢.

Thus, player i knows that any hypothetical event where he chooses differently than what he actually observed himself choosing will be assigned zero probability by all the other players. Moreover, player i also assigns zero probability to any such hypothetical event, because he also knows any two distinguishable actions $a_i$ ¹ $a_i$¢ represent disjoint events that never happen simultaneously.

8. The last conclusion of step 7 implies that player i knows that the true state of the world w Î W must have hypothetically changed whenever he hypothetically contemplates changing his action from any given action $a_i$ Î $A_i$; because no single w can simultaneously be contained in two disjoint subsets of W whose intersection is empty [$a_i$ Ç $a_i$¢ = Æ]. Consequently, Aumann's definition (1987, page 7) of player i being "Bayes rational ***at*** [a single] w" either does not make sense conceptually; or at least is an unfruitful way of defining Bayesian rationality when players know nothing more than their observed actions. With such limited knowledge, Bayesian rationality necessarily requires *different* states of the world to be hypothetically compared when player i compares his expected payoffs from hypothetically choosing different actions *[because hypothetically having chosen different actions implies different states must have hypothetically occurred*]. This does not mean player i ever knows exactly what specific state w Î W exists when he chooses; just that w must be different than what it would have been had he chosen differently.

9. The latter conclusion of step 8 implies the following inference: if a player's information equals his observed action and he contemplates hypothetically choosing differently; then he necessarily also contemplates different information (about which specific state w Î W exists) if he chooses differently; which in turn implies the relative likelihood of other player's actions conditional on his different information may thereby also be different if he chooses differently [*choosing differently implies the true state of the world must be different; which implies the likelihood of other players' actions may also be different for a different state of the world*]. Consequently, player i cannot assume in such a situation that the probabilities of other players' actions (conditional on his choosing any given action) are the same if he contemplates hypothetically deviating from choosing that action [that is, player i cannot assume that the probabilities of other players' actions are statistically independent of changes in his own hypothetically chosen action].

Yet, such an independence is what Aumann (1987) implicitly assumes in his formal theorem characterizing a correlated equilibrium distribution (proposition 2.3 on page 6). This is because the conditional probabilities of other players' choices (denoted $p_{jk}/S_k p_{jk}$ in Aumann's notation) are held fixed as a player hypothetically changes its actions (in Aumann's proof of proposition 2.3). Consequently, Aumann's formal characterization of a correlated equilibrium is inconsistent with his other assumptions: that each player knows what action he chooses; and that states of the world are "comprehensively defined" to include his own choice (along with everyone else's choices).

10. A simple way to avoid the preceding inconsistency is to recognize that Aumann's two assumptions (noted just above) do not require that players' subjective prior distributions $s_i$ imply "exogenous" information sources, as described in Definition 16 of the main text. Instead these prior distributions are consistent with both "exogenous" and "endogenous" information sources; where the latter allows the conditional probabilities between players' messages p($m_i$½$m_j$) to depend on their profile of actual choices. That is, each player's prior distribution is also consistent with $s_i$($m_i$, $m_j$½a) ¹ $s_i$($m_i$,$m_j$½a¢) for some a ¹ a¢ Î A; where $s_i$($m_i$, $m_j$½a) equals the ratio of $s_i$($m_i$, $m_j$, a) divided by $s_i$(a), and similarly for $s_i$($m_i$, $m_j$½a¢). In short, assuming Bayesian players' (who have prior distributions $s_i$ over potential states w Î W) does not presuppose whether their prior beliefs allow exogenous or endogenous information sources.

**REFERENCES**

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

Axelrod, Robert and Douglas Dion. 1988. "The Further Evolution of Cooperation," *Science*, **242** (December 9), pp. 1385-1390.

Aumunn, Robert.1987. "Correlated Equilibria as an Extension of Bayesian Rationality," *Econometrica*, **55**(1), January, pp. 1-18.

Aumunn, Robert. 1974. "Subjectivity and Correlation in Randomized Strategies," *Journal of Mathematical Economics*, **1**, pp. 67-96.

Binmore, Ken. 1992. *Fun and Games*, Toronto, Heath.

Binmore, Ken. 1994. *Game Theory and the Social Contract I: Playing Fair*, MIT Press, Cambridge.

Binmore, Ken, Avner Shaked, and John Sutton. 1985. "Testing Noncooperative Bargaining Theory: A Preliminary Study," *American Economic Review*, **75**, pp. 1178-1180.

Bolle, F. and p. Ockenfels. 1990. "Prisoner's Dilemma as a Game with Incomplete Information," *Journal of Economic Psychology*, **11** p. 69 ff.

Bonner, John T. 1980. *The Evolution of Culture in Animals*, Princeton, Princeton University Press.

Brandenberger, Adam. 1992. "Knowledge and Equilibrium in Games," *Journal of Economic Perspectives*, **6**(4), pp. 83-101.

Burgoon, Michael, Frank Hunsaker and Edwin Dawson. 1994. *Human Communication* (3rd ed.), Thousand Oaks, Sage.

Coase, Ronald. 1937. "The Nature of the Firm," *Economica*, **4**, PP. 386-405.

Coleman, James S. 1990. *Foundations of Social Theory*, Cambridge, Harvard University Press.

Cornell, Stephen and Joseph P. Kalt. 1995a. "Where Does Economic Development Really Come From? Constitutional Rule Among the Contemporary Sioux and Apache," *Economic Inquiry*, forthcoming.

Cornell, Stephen and Joseph P. Kalt. 1995b. "Cultural Evolution and Constitutional Public Choice: Institutional Diversity and Economic Performance on American Indian Reservations," research paper R95-8, John F. Kennedy School of Government, Harvard University.

Cornell, Stephen and Joseph P. Kalt. 1995c. "Successful Economic Development and Heterogeneity of Governmental Form on American Indian Reservations," research paper, March, John F. Kennedy School of Government, Harvard University.

Crawford, Vincent. 1990. "Explicit Communication and Bargaining Outcomes," *American Economic Review*, **80**, pp. 213-219.

Davies, D. R. 1969. H*uman Vigilance Performance,* New York, Elsevier.

Dawes, Robyn. 1980. "Social Dilemmas," *Annual Review of Psychology*, **31**, pp. 169-193.

Dawes, Robyn, David Faust, and Paul Meehl. 1989. "Clinical Versus Actuarial Judgment," S*cience*, **243** (March 31), pp. 1668-1674.

Dawkins, Richard. 1976. *The Selfish Gene*, Oxford, Oxford University Press.

Egan, James. 1975. *Signal Detection Theory and ROC Analysis*, New York, Academic Press.

Farrell, Joseph. 1987. "Cheap Talk, Coordination, and Entry," *Rand Journal of Economics*, **18**, pp. 34-39.

Fishburn, Peter. 1982. *Foundations of Expected Utility*, New York, Cambridge University Press.

Fudenberg, Drew and Jean Tirole. 1991. *Game Theory*, Cambridge, MIT.

Frank, Robert. 1988. *Passions Within Reason: Prison's Dilemmas and the Strategic Role of Emotions*, New York, W. W. Norton.

Frank, Robert, R. Gilovich, & R. Regan. 1993. "The Evolution of One-Shot Cooperation: An Experiment," *Ethology and Sociobiology* , **14**, pp. 247-256.

Frey, Bruno and Iris Bohnet. 1994. "Cooperation and Fairness in Experiments: Relevance for Democracy," *manuscript*, Institute for Empirical Economic Research, University of Zurich, Switzerland.

Gauthier, David. 1986. *Morals By Agreement*, Oxford, Oxford University Press.

Geanakoplos, John. 1992. "Common Knowledge," *Journal of Economic Perspectives*, **6**(4), pp. 53-82.

Gibbard, Alan, and William Harper. 1978. "Counterfactuals and Two Kinds of Expected Utility," in *Foundations and Applications of Decision Theory*, Hooker, Leach, and McClennen, ed., Reidel Publishing.

Green, David and John Swets. 1974. *Signal Detection Theory and Psychophysics*, New York, Robert Kriegur.

Guth, Werner and Hartmut Kliemt. 1994. "Competition or Cooperation: On the Evolutionary Economics of Trust, Exploitation and Moral Attitudes," *Metroeconomica*, **45** (2). pp. 155-187.

Guth, Werner and Reinhard Tietz. 1990. "Ultimatum Bargaining Behavior: A Survey and Comparison with Experimental Results," *Journal of Economic Psychology*, **11**, pp. 417-449.

Harsanyi, J. C. 1967. "Games with Incomplete Information Played by Bayesian Players," *Management Science*, **14**, (parts I, II, III), pp. 159-182, 320-334, 486-502.

Harsanyi, J. C. 1977. *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*, Cambridge, Cambridge University Press.

Heiner, Ronald A. 1983. "The Origin of Predictable Behavior," *American Economic Review*, **73**, pp. 560-595.

Heiner, Ronald A. 1986a. "Uncertainty, Signal Detection Experiments, and Modeling Behavior," in *Economics as a Process: Essays in the New Institutional Economics*, R. Languois, ed., New York, Cambridge University.

Heiner, Ronald A. 1986b. "Imperfect Decisions and the Law: On the Evolution of Legal Precedent and Rules," *Journal of Legal Studies*, **15**, pp. 227-261.

Heiner, Ronald A. 1988a. "The Necessity of Imperfect Decisions," *Journal of Economic Behavior and Organization*, **10**, pp. 29-56.

Heiner, Ronald A. 1991. "Origin of Predictable Dynamic Behavior," *Journal of Economic Behavior and Organization*, **12**, pp. 233-258.

Isaac, Mark and James Walker. 1988. "Communication and Free-Riding Behavior: The Voluntary Contribution Mechanism," *Economic Inquiry*, **24**, pp. 585-608.

Johnson, James. 1993. "Is Talk Really Cheap? Prompting Conversation Between Critical Theory and Rational Choice," *American Political Science Review*, **87**, pp. 74-86.

Alchian, Armen. 1984. "Specificity, Specialization, and Coalitions," *Journal of Institutional & Theoretical Economics*, **140**, pp. 34-49.

Klein, B., Crawford, R., & Alchian, A. 1978. "Verticle Integration, Appropriable Rents, and the Competitive Contracting Process," *Journal of Law & Economics*, **21**, pp. 297-326.

Kreps, David and Robert Wilson. 1982. "Sequential Equilibria," *Econometrica*, **50**, pp. 863-894.

Liebrand, W., David Messick, and Henk Wilke. 1992 *Social Dilemmas: Theoretical Issues and Research Findings*, Permagon Press, New York.

Machina, Mark. 1985. "Nonexpected Utility Theory: A survey," Institute for Mathematical Economics, Stanford University.

McNicol, D. 1972. *A Primer on Signal Detection Theory*, London, Allen & Urwin.

Nevin, John. 1969. "Signal Detection Theory and Operant Behavior," *Journal of the Experimental Analysis of Behavior*, **12**, pp. 475-480.

Nisbett, Robert and Tom Wilson. 1977. "Telling More Than We Can Know: Verbal Reports on Mental Processes," *Psychological Review*, **84**(3), pp. 231-259.

Nozick, Robert. 1969. "Newcomb's Problem and Two Principles of Choice," in *Essays in Honor of Carl G. Hempel*, Nicholas Rescher, ed., Reidel, Dortrecht-Holland.

Orbell, John, Alphons van de Kragt and Robyn Dawes. 1988. "Explaining Discussion-Induced Cooperation," *Journal of Personality and Social Psychology*, **54**, pp. 811-819.

Ostrom, Elinor. 1990. *Governing the Commons*, Cambridge University Press, New York.

Powell, Walter. 1990. "Neither Markets or Hierarchy: Network Forms of Organization," *Research in Organizational Behavior*, **12**, pp. 295-336.

Reny, Philip. 1992. "Rationality in Extensive Form Games," *Journal of Economic Perspectives*, **6**(4), pp. 103-118.

Rosenthal, R. W. 1981. "Games of Perfect Information, Predatory Pricing and the Chain Store Paradox," *Journal of Economic Theory*, **25**, p. 92 ff.

Schelling, Thomas. 1960. *The Strategy of Conflict*, Harvard University Press, Cambridge.

Schultz, Ulrich, Wulf Albers, & Ulrich Mueller. 1993. *Social Dilemmas and Cooperation*, Springer-Verlag, New York.

Selton, R. 1988. "Evolutionary Stability in Extensive Two Person Games," *Mathematical Social Science*, **16**, p. 223 ff.

Smith, John. 1977. *The Behavior of Communicating*, Harvard, Harvard University Press.

Staddon, John. 1983. *Adaptive Behavior and Learning*, Cambridge, Cambridge University Press.

Sugden, Robert. 1995. "A Theory of Focal Points," *The Economic Journal*, **105**(430) May, pp. 533-550.

Swets, John (ed.). 1964. *Signal Detection and Recognition by Human Observers*, New York, Wiley.

Swets, John and R. M. Pickett. 1982. *Evaluation of Diagnostic Systems: Methods From Signal Detection Theory*, New York, Academic Press.

Swets, John. 1988. "Measuring the Accuracy of Diagnostic Systems," *Science*, **240** (June 3rd), pp. 1285-1293.

Tanner, T., J. Rauk, and R. Atkinson. 1970. "Signal Recognition as Influenced by Information Feedback," *Journal of Mathematical Psychology*, **7**, pp. 259-274.

Williamson, Oliver. 1975. *Markets and Hierarchies*, New York, Cambridge University.

Wilson, Edward O. 1975. *Sociobiology*, Harvard, Harvard University Press (Belknap).