Supplement to Foundations for Cooperation in the Prisoners' Dilemma

Brendan Daley

Duke University

Philipp Sadowski

Duke University

May 7, 2014

ERID Working Paper Number 155

Economic Research Initiatives at Duke
WORKING PAPERS SERIES

Supplement to
"Foundations for Cooperation in the Prisoners' Dilemma"
(For Online Publication)

Brendan Daley[†] Philipp Sadowski[‡]

May 7, 2014

Abstract

We establish that in the Prisoners' Dilemma, the model of Daley and Sadowski (2014) is logically distinct from three models that employ well-known forms of other-regarding preferences: altruism (Ledyard, 1995; Levine, 1998), inequity aversion (Fehr and Schmidt, 1999), and reciprocity (Rabin, 1993).

S.1 Introduction

In this supplement we consider the class of prisoners' dilemma games denoted PD from Section 3 of Daley and Sadowski (2014) (henceforth DS14) in which each game is parameterized by a pair $(x, y) \in R^2_{++}$.[1] Each player i in a finite collection I reports his preferred action for each possible game in PD as if it were to be played as a one-shot game against an anonymous opponent. D_i, C_i, and M_i denote the games in which i strictly prefers defection, cooperation, or is indifferent, respectively.

The behavioral model in DS14 is as follows. Each player $i \in I$ is privately-endowed with a type α_i. In addition, there is a common prior that types are drawn i.i.d. from an atomless distribution with support $[0, 1]$ and differentiable CDF denoted F. For each $(x, y) \in PD$, in any equilibrium (σ, P), player i evaluates the expected payoff of action $a_i \in \{c, d\}$ as:

$$V_i(c|x, y, P) = \alpha_i \cdot 1 + (1 - \alpha_i)[P \cdot (-y) + (1 - P) \cdot 1]$$
$$V_i(d|x, y, P) = \alpha_i \cdot 0 + (1 - \alpha_i)[P \cdot 0 + (1 - P)(1 + x)]$$

where α_i is interpreted as i's degree of magical thinking.

The representation result establishes that under a condition on the slope of F, the data generated by the unique equilibrium behavior in PD of any such model satisfies four axioms, and for any data set that satisfies the axioms there exists a model, satisfying the same slope condition on F, that can explain it.[2]

Of course, there may exist other equivalent representations. As well-known models employing what are referred to as "other-regarding preferences" can sometimes accommodate

[†]The Fuqua School of Business, Duke University. E-mail: bd28@duke.edu
[‡]Department of Economics, Duke University. E-mail: p.sadowski@duke.edu

[1]Because the purpose of this supplement is to demonstrate that the alternative models are behaviorally distinct from that of DS14, it suffices to establish the result on the subclass of games $PD \subset PD^0$.

[2]Recall that the four axioms are Axioms 2-5, as Axiom 1 is needed only when considering the larger class of games PD^0.

cooperation by some players in some games in PD, they may seem candidates for this equivalence. In this supplement, we demonstrate that the models endowed with three of the most popular forms of other-regarding preferences are logically distinct from our model on our domain.[3]

Let u_i, u_j be the payoffs to players i and j as specified by the outcome of a two-player game. In each of the three models, player i seeks to maximize a different objective, which we denote v_i.

1. **Altruism**. As proposed by Ledyard (1995) and further studied by Levine (1998): $v_i = u_i + \alpha_i u_j$, where $0 \leq \alpha_i < 1$; player i may care about his opponent's payoff, but not more than his own.

2. **Inequity Aversion**. As proposed by Fehr and Schmidt (1999):[4] $v_i = u_i - \alpha_i \max\{u_j - u_i, 0\} - \delta\alpha_i \max\{u_i - u_j, 0\}$, where $0 \leq \alpha_i$ and $0 < \delta < \min\left\{1, \frac{1}{\alpha_i}\right\}$; player i may dislike inequity, but dislikes it more if his is the smaller payoff, and is not willing to "burn" his own payoff to create equity.

3. **Reciprocity**. As proposed by Rabin (1993), player i cares about how "fair" he and his opponent are being to one another. Fixing the action of player i, a_i, how fair player j is being to player i is captured by the "kindness" function $K_j(a_j|a_i)$. In the prisoners' dilemma, once a_i is fixed all outcomes are Pareto optimal. In this case,

$$K_j(a_j|a_i) = \frac{u_i(a_i, a_j) - \frac{1}{2}\left(u_i^h(a_i) + u_i^l(a_i)\right)}{u_i^h(a_i) - u_i^l(a_i)}$$

where $u_i^h(a_i)$ and $u_i^l(a_i)$ are, respectively, the highest and lowest possible payoffs to i given a_i. Finally, $v_i = u_i + \alpha_i K_j(1 + K_i)$, where $\alpha_i \geq 0$.

S.2 Results

The original specifications of these models did not include heterogeneity in the degree to which players are other-regarding. To incorporate heterogeneity into these models, in each we assume there is a common prior that α-types are drawn i.i.d. from a continuous distribution with support $[\underline{\alpha}, \overline{\alpha}]$, where $\underline{\alpha} \geq 0$, and CDF F. Complete homogeneity can be thought of as a limiting case as $(\overline{\alpha} - \underline{\alpha}) \to 0$. Our equilibrium notion is just as in Definition 3.2 of DS14, with $V_i(\cdot)$ suitably adapted to each model.

It is not our goal here to provide a comprehensive analysis of these models (which, while doable, would require a considerably longer treatment), but to establish the following.

[3] DS14 discusses how magical thinkers differ from other-regarding players beyond their play in PD. In addition, as discussed in Section 2.1 of DS14, altruism and inequity aversion can only explain cooperation under the *physical* interpretation of payoffs, which we therefore adopt in this supplement.

[4] One could consider a more general version in which the $\delta\alpha_i$ term is replaced by β_i. That is, players can have two-dimensional types. This would not alter our result.

Proposition S.1 *Fix any model of those described above and an equilibrium, (σ_g, P_g), for each game $g \in PD$ and consider the resultant data of all collections of arbitrary size n. Either, for all collections I, $D_i = PD$ for all $i \in I$, or there is a positive measure of collections (according to the common prior, F) each of whose data violates Axioms 2-5 of DS14.*[5]

The result is proved in the subsequent analysis.

S.2.1 Altruism

Fixing any $(x, y) \in PD$ and an equilibrium (σ, P),

$$\begin{aligned} V_i(c|x,y,P) &= (1-P)(1+\alpha_i \cdot 1) + P(-y + \alpha_i(1+x)) \\ V_i(d|x,y,P) &= (1-P)(1+x+\alpha_i(-y)) + P(0+\alpha_i \cdot 0) \end{aligned}$$

Therefore, $V_i(c|x,y,P) - V_i(d|x,y,P) = \alpha_i(1+Px+(1-P)y) - (1-P)x - Py$. This expression is strictly increasing in α_i for all $(x, y), P$. Hence, all equilibria are cutoff equilibria.

For any given $(x, y) \in PD$, there exists an equilibrium with cutoff type α if and only if given $\alpha_i = \alpha$, $V_i(c|x, y, F(\alpha)) = V_i(d|x, y, F(\alpha))$. For any α, let $\tilde{M}_\alpha$ be the set of games in which there exists an equilibrium in which α is the cutoff type. Algebraically,

$$\tilde{M}_\alpha = \left\{ (x,y) \in PD \middle| y = \frac{\alpha}{(1+\alpha)F(\alpha) - \alpha} - \frac{1-(1+\alpha)F(\alpha)}{(1+\alpha)F(\alpha)-\alpha} \cdot x \right\}$$

Clearly, for all $i \in I$, $M_i \subset \tilde{M}_{\alpha_i}$.

We now argue that for any F, there exists a (generic) collection drawn from its support whose equilibrium play violates the axioms. First, let $\alpha^0 < \overline{\alpha}$ be the unique solution to $F(\alpha^0) = \frac{1}{1+\alpha^0}$. For all $\alpha \in [\alpha^0, \overline{\alpha}]$, $\tilde{M}_\alpha$ forms a line in PD that is weakly *upward* sloping. So, for any player i with $\alpha_i \in [\alpha^0, \overline{\alpha}]$ to be consistent with *Continuity* (Axiom 2) and *Monotonicity* (Axiom 3), it must be that $M_i = \emptyset$.[6] Second, fix arbitrary $\alpha \in [\alpha^0, \overline{\alpha}]$. Simple algebra shows that in the game $\left(\frac{\alpha}{1-\alpha}, \frac{\alpha}{1-\alpha}\right) \in PD$, α is the unique equilibrium cutoff, so must be in M_i for any i such that $\alpha_i = \alpha$. Hence, any player drawn from a high enough quantile of the distribution will have a violation.

The intuition for this is easy to see. Suppose that $\alpha_i = \overline{\alpha}$, so $F(\alpha_i) = 1$. Then, if in game (x, y), i is indifferent between c and d, all other players are choosing d. Therefore, i's indifference condition is $V_i(c|x, y, 1) = -y + \alpha_i(1 + x) = V_i(d|x, y, 1) = 0$. An increase in x *increases* $V_i(c)$ because it increases i's opponent's payoff, which i values altruistically. This makes player i strictly prefer c to d, and violates *Monotonicity*.

[5] Further, the proposition remains valid if collections are formed via i.i.d. draws from any distribution with support $[\underline{\alpha}, \overline{\alpha}]$, even if its CDF differs from the one perceived by the players, F.

[6] Suppose not, and that $(x, y) \in M_i$. Then to satisfy Axiom 3, i) all other $(x', y') \in \tilde{M}_{\alpha_i}$ cannot be in M_i (so $M_i = \{(x, y)\}$), and ii) $C_i \neq \emptyset$ and $D_i \neq \emptyset$. But then Axiom 2 is clearly violated.

S.2.2 Inequity Aversion

Fixing any $(x,y) \in PD$ and an equilibrium (σ, P),

$$V_i(c|x,y,P) = (1-P)(1-\alpha_i \cdot 0) + P(-y-\alpha_i(1+x+y))$$
$$V_i(d|x,y,P) = (1-P)(1+x-\delta\alpha_i(1+x+y)) + P(0-\alpha_i \cdot 0)$$

Therefore, $V_i(c|x,y,P) - V_i(d|x,y,P) = \alpha_i(1+x+y)(\delta - P(1+\delta)) - (1-P)x - Py$. This expression is negative for $\alpha_i = 0$, monotonic in α_i, and increasing in α_i if and only if $P < \frac{\delta}{1+\delta} \leq \frac{1}{2}$. This immediately implies that all players defecting regardless of type (i.e., $P = 1$) is an equilibrium for any $(x,y) \in PD$. It also implies that if, for a given game, there exists an equilibrium in which a type cooperates, then it is a cutoff equilibrium where the cutoff type α^* must satisfy $F(\alpha^*) < \frac{\delta}{1+\delta} \leq \frac{1}{2}$.

Fix now any player i with α_i such that $F(\alpha_i) > \frac{1}{2}$. From above, $M_i = \emptyset$. Notice, though, that in any game, in any equilibrium where any type cooperates, player i cooperates. Therefore, we have the following two cases:

Case 1: Suppose $C_i = \emptyset$. Then, by the previous paragraph, in every game players are coordinating on the "all defect" equilibrium. Therefore, $D_j = PD$ for all $j \in I$, consistent with Proposition S.1.

Case 2: Suppose $C_i \neq \emptyset$. Then, given $M_i = \emptyset$, for player i to satisfy *Continuity* (Axiom 2), it must be that $D_i = \emptyset$. We now show that this cannot hold. To see this notice that i) $V_i(c|x,y,P) - V_i(d|x,y,P)$ is monotonic (in fact, linear) in P, and ii) $V_i(c|x,y,1) - V_i(d|x,y,1) = -y - \alpha_i(1+x+y) < 0$ for all α_i and $(x,y) \in PD$. Therefore, if $V_i(c|x,y,0) - V_i(d|x,y,0) < 0$, then there is no equilibrium for game (x,y) in which i cooperates.

$$V_i(c|x,y,0) - V_i(d|x,y,0) = \delta\alpha_i(1+y) + x(-1+\delta\alpha_i)$$

Since $\delta\alpha_i < 1$, this is negative if $x > \frac{\delta\alpha_i(1+y)}{1-\delta\alpha_i}$. For any fixed y, there exist large enough x-values to satisfy this inequality for all α_i. Hence, $D_i \neq \emptyset$, violating Axiom 2.

S.2.3 Reciprocity

It is easy to calculate that for any pair of players i, j and $(x,y) \in PD$, regardless of a_i, $K_j(a_j = d|a_i) = -\frac{1}{2}$ and $K_j(a_j = c|a_i) = \frac{1}{2}$. So, fixing any $(x,y) \in PD$ and an equilibrium (σ, P),

$$V_i(c|x,y,P) = (1-P)(1+\tfrac{3}{4}\alpha_i) + P(-y-\tfrac{3}{4}\alpha_i)$$
$$V_i(d|x,y,P) = (1-P)(1+x+\tfrac{1}{4}\alpha_i) + P(0-\tfrac{1}{4}\alpha_i)$$

From here, the analysis is analogous to that performed for inequity-averse players. $V_i(c|x,y,P) - V_i(d|x,y,P) = \alpha_i\left(\frac{1}{2} - P\right) - (1-P)x - Py$. This expression is negative for $\alpha_i = 0$, monotonic in α_i, and increasing in α_i if and only if $P < \frac{1}{2}$. This immediately implies that all players defecting regardless of type (i.e., $P = 1$) is an equilibrium for any $(x, y) \in PD$. It also implies that if, for a given game, there exists an equilibrium in which a type cooperates, then it is a cutoff equilibrium where the cutoff type α^* must satisfy $F(\alpha^*) < \frac{1}{2}$.

Fix now any player i with α_i such that $F(\alpha_i) > \frac{1}{2}$. From above, $M_i = \emptyset$. Notice, though, that in any game, in any equilibrium where any type cooperates, player i cooperates. Therefore, we have the following two cases:

Case 1: Suppose $C_i = \emptyset$. Then, by the previous paragraph, in every game players are coordinating on the "all defect" equilibrium. Therefore, $D_j = PD$ for all $j \in I$, consistent with Proposition S.1.

Case 2: Suppose $C_i \neq \emptyset$. Then, given $M_i = \emptyset$, for player i to satisfy *Continuity* (Axiom 2), it must be that $D_i = \emptyset$. We now show that this cannot hold. To see this notice that i) $V_i(c|x, y, P) - V_i(d|x, y, P)$ is monotonic (in fact, linear) in P, and ii) $V_i(c|x, y, 1) - V_i(d|x, y, 1) = -\left(\frac{\alpha_i}{2} + y\right) < 0$ for all α_i and $(x, y) \in PD$. Therefore, if $V_i(c|x, y, 0) - V_i(d|x, y, 0) < 0$, then there is no equilibrium for game (x, y) in which i cooperates.

$$V_i(c|x, y, 0) - V_i(d|x, y, 0) = \frac{\alpha_i}{2} - x$$

This is negative if $x > \frac{\alpha_i}{2}$. Hence, $D_i \neq \emptyset$, violating Axiom 2.

References

Daley, B. and P. Sadowski (2014): "Foundations for Cooperation in the Prisoners' Dilemma," *Working Paper*.

Fehr, E. and K. Schmidt (1999): "A Theory of Fairness, Competition, and Cooperation," *The Quarterly Journal of Economics*, 114, 817–868.

Ledyard, J. (1995): "Public Goods: A Survey of Experimental Research," in *Handbook of Experimental Economics*, ed. by J. H. Kagel and A. E. Roth.

Levine, D. K. (1998): "Modeling Altruism and Spitefulness in Experiments," *Review of Economic Dynamics*, 1(3), 593–622.

Rabin, M. (1993): "Incorporating Fairness into Game Theory and Economics," *The American Economic Review*, 83(5), 1281–1302.