Journal of Mathematical Psychology 67 (2015) 45–50

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

journal homepage: www.elsevier.com/locate/jmp

Social preferences, positive expectations, and trust based cooperation☆

Ryan O. Murphy*, Kurt A. Ackermann

ETH Zürich, Chair of Decision Theory and Behavioral Game Theory, Clausiusstrasse 50, 8092 Zürich, Switzerland

HIGHLIGHTS

- We develop a psychologically grounded model of trust based cooperation.
- We integrate SVO, beliefs, trust, and cooperation among interdependent players.
- Trust thresholds can be derived over combinations of social preferences and beliefs.
- Rapoport's K-index is the minimum SVO to justify cooperation given a uniform prior.
- Different joint utility functions affect when trust based cooperation is expected.

ARTICLE INFO

Article history:
Received 31 October 2014
Received in revised form
11 June 2015
Available online 10 July 2015

Keywords:
Trust
Cooperation
Social preferences
Beliefs
Prisoner's dilemma
Index of cooperation
Rationalizability
SVO

ABSTRACT

Some accounts of cooperation in the Prisoner's Dilemma have focused on developing simple indexes of a game's *severity* – i.e., the degree to which a game promotes non-cooperative choices – which are derived wholly from the game's payoff structure. However, the psychological mechanisms of *why* a game's payoffs affect cooperation rates are not clearly explicated with this approach. We show how simple models of decision making can predict the emergence of trust based cooperation as the expected utility maximizing strategy when individual social preferences and positive expectations (beliefs) are simultaneously taken into account. Moreover, we show how these predictions relate to a particular game's index of cooperation. We then delineate under what conditions trust based cooperation is rationalizable, and how the decision to trust can be understood in terms of an interaction between payoffs, preferences, and beliefs.

1. Introduction

It has been shown empirically that cooperation rates are systematically associated with Prisoner Dilemma (PD) games' payoff structures (Glöckner & Hilbig, 2012; Rapoport & Chammah, 1965; Steele & Tedeschi, 1967; Vlaev & Chater, 2006), which has led researchers to devise metrics for predicting the aggregate cooperation rates from the payoff structures of games (see Fig. 1 for the PD game). Arguably the best-known metric of this kind is Rapoport's (1967) K-index of cooperation, however there are also others (e.g., Axelrod, 1967; Roth & Murnighan, 1978). Rapoport's index is based on two considerations: namely that (i) the higher the payoffs potentially resulting from cooperation (i.e., R and S), the higher the expected cooperation rate; and that (ii) the higher the payoffs potentially resulting from defection (i.e., T and P), the lower the expected cooperation rate. The K-index incorporates these two factors by dividing the difference between the best payoff a decision maker (DM) can receive from cooperating and the worst payoff the DM can receive from defecting, by the difference between the best payoff from defecting and the worst payoff from cooperating: $\frac{(R-P)}{(T-S)}$. Hence, the K-index captures, at least to some extent, the severity (we use the term *severity* consistent with its definition by Rapoport and Chammah (1965) to refer to the general temptation to defect) of the Prisoner's Dilemma game. The higher the K-index, the *less* severe is the dilemma, and thus higher rates of cooperation are anticipated, all other things being equal.

But the severity of a PD game as a function of its payoffs can only have an effect on DMs' behavior if the DMs have positive other-regarding preferences (i.e., a DM derives some positive utility from the other player's payoff). Furthermore, given a PD game's particular payoff structure, and a DM's particular degree of concern for the other player's payoff, the choice to cooperate will also depend in

☆ This research has been supported in part by Swiss National Science Foundation (SNF) grant 100014_143199/1.

* Corresponding author.
E-mail address: rmurphy@ethz.ch (R.O. Murphy).

http://dx.doi.org/10.1016/j.jmp.2015.06.001
0022-2496/

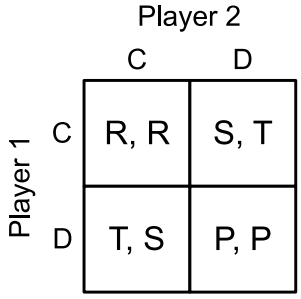

Fig. 1. The standard Prisoner's Dilemma game.

part on the DM's belief about whether the other player will choose to cooperate as well. That is, the underlying determinants of cooperation in the PD game are preferences and beliefs, within the context of a particular payoff structure. In order to make precise predictions about cooperation rates in a PD game, the interplay between these three factors has to be taken into account.

This last statement provides the central point for this paper. We use a simple model of a DM's social preferences, and their beliefs about the other players' anticipated choice, and we use these two factors simultaneously to predict when a DM will choose to trust and thus act cooperatively in a one-shot PD game, given the game's particular payoff structure. Furthermore, we show how different indexes of cooperation can be extracted from such models, and how they relate to the K-index of cooperation and each other. Although these summary indexes are useful, the psychological factors that are responsible for trust based cooperation are of primary interest.

2. Elements

2.1. The Prisoner's Dilemma (PD) game

In this paper we consider standard 2×2 symmetric PD games (see Fig. 1). The Prisoner's Dilemma game in normal form is instantiated when the payoffs conform to the strict inequalities $T > R > P > S$. Although not a necessary characteristic, we also limit our consideration to games where $2R > (T + S)$. To focus attention, let us anchor $T = 1$ and $S = 0$ for all the games. Further, let us restrict R and P to be evenly divisible by 0.1. This reduces the number of PD games we will consider but does so without any loss of generality and evenly covers the space of possible PD games.

This discrete configuration yields 26 different PD games. The games are shown in Table 1 with each of the games' corresponding K-index, as well as other summary indexes which are explained in more detail later in the paper. Note that different PD games can have the same K-index.

2.2. Social preferences

There is ample evidence that people are heterogeneous in the way they evaluate joint payoffs (Van Lange, 1999), and that other-regarding preferences can be rationalized in a utility framework (e.g., Andreoni & Miller, 2002). The most basic representation of social preferences can be implemented with a joint utility function for a decision maker that attaches a single parameter (α) to the other player's payoff:

$$u(\pi_s, \pi_o) = \pi_s + \alpha \cdot \pi_o. \quad (1)$$

Here π_s is the DM's payoff (the payoff for the self), and π_o is the other player's payoff. Alpha is an index of other-regarding preferences and is consistent with the concept of Social Value Orientation (for reviews on SVO see Au & Kwong, 2004; Murphy & Ackermann, 2014). Narrow self-interest can be accommodated in this framework when α equals zero.

Table 1
These are all possible PD games with $T = 1$, $S = 0$, R and P in steps of 0.1, and conforming to the inequalities in Section 2.1.

PD game	T	R	P	S	K	CoopArea	α_{crit}	PoA
1	1	0.6	0.5	0	0.10	0.18	0.82	1.20
2	1	0.7	0.6	0	0.10	0.21	0.82	1.17
3	1	0.8	0.7	0	0.10	0.26	0.82	1.14
4	1	0.9	0.8	0	0.10	0.30	0.82	1.13
5	1	0.6	0.4	0	0.20	0.33	0.67	1.50
6	1	0.7	0.5	0	0.20	0.32	0.67	1.40
7	1	0.8	0.6	0	0.20	0.33	0.67	1.33
8	1	0.9	0.7	0	0.20	0.35	0.67	1.29
9	1	0.6	0.3	0	0.30	0.46	0.54	2.00
10	1	0.7	0.4	0	0.30	0.46	0.54	1.75
11	1	0.8	0.5	0	0.30	0.43	0.54	1.60
12	1	0.9	0.6	0	0.30	0.42	0.54	1.50
13	1	0.6	0.2	0	0.40	0.56	0.43	3.00
14	1	0.7	0.3	0	0.40	0.57	0.43	2.33
15	1	0.8	0.4	0	0.40	0.56	0.43	2.00
16	1	0.9	0.5	0	0.40	0.53	0.43	1.80
17	1	0.6	0.1	0	0.50	0.65	0.33	6.00
18	1	0.7	0.2	0	0.50	0.66	0.33	3.50
19	1	0.8	0.3	0	0.50	0.66	0.33	2.67
20	1	0.9	0.4	0	0.50	0.65	0.33	2.25
21	1	0.7	0.1	0	0.60	0.74	0.25	7.00
22	1	0.8	0.2	0	0.60	0.75	0.25	4.00
23	1	0.9	0.3	0	0.60	0.74	0.25	3.00
24	1	0.8	0.1	0	0.70	0.82	0.18	8.00
25	1	0.9	0.2	0	0.70	0.82	0.18	4.50
26	1	0.9	0.1	0	0.80	0.89	0.11	9.00

2.3. Beliefs—positive expectations of the other player

Here we posit that a DM believes that the other player will choose strategy C with a probability of β. If the DM is certain that the other player will cooperate, then β equals 1; conversely if the DM is certain the other player will defect, then β equals 0. Gradations between these two extremes are captured by different β values in the probability space from [0, 1]. The standard normative model posits that DMs believe with certainty that other players will never choose strategy C. Models where DMs may have some non-zero expectation of the other player have been previously developed; perhaps the best known work along this line is Kreps, Milgrom, Roberts, and Wilson (1982).

2.4. Trust

We contend that a DM choosing to cooperate in a PD game is manifesting trust, as the PD game is fundamentally a kind of simple trust game; more specifically the PD is a two-player, two-option, symmetric, simultaneous, trust game (*cf.* Berg, Dickhaut, & McCabe, 1995). Along these lines, the choice to cooperate demonstrates both positive intentions and positive expectations on behalf of the DM. This viewpoint is consistent with well-known definitions for trust. Take for instance Rousseau, Sitkin, Burt, and Camerer (1998, p. 395): "Trust is a psychological state comprising the intention to accept vulnerability based upon positive expectations of the intentions or behavior of another". We suggest an addendum to this definition as follows: "…with the intention of improving collective outcomes". This is a useful addition in that it highlights that trust is an intentional choice and that when choosing to cooperate, a DM has some prosocial preferences and a goal in mind, namely to promote collective efficiency which is valued by the DM. Moreover this addendum to the definition offers an explanation of *why* a DM would volunteer to take on the strategic risk of being exploited by the other player. The reason in our view is that the