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

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/

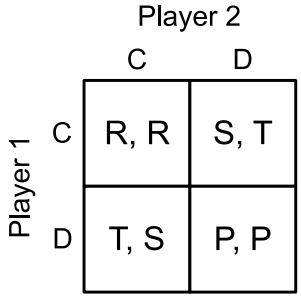

Fig. 1. The standard Prisoner's Dilemma game.

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

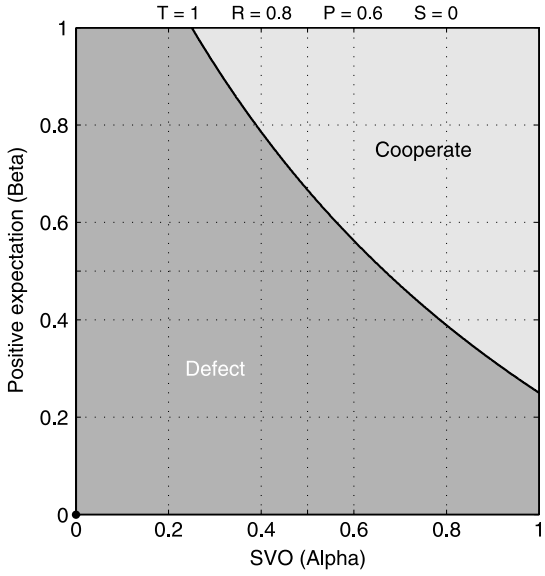

Fig. 2. This plot shows when trust based cooperation would be expected in a PD game with payoffs $T = 1, R = 0.8, P = 0.6, S = 0$. Different SVOs and different positive expectations of cooperation from the other player are represented on the x and y axes. The dark area indicates when *defect* is a best response, whereas the light area indicates when *cooperate* is a best response. As expected, DMs with both high SVO and high positive expectations would choose to cooperate. The idealized DM *homo economicus* is just a special case and is located at the origin, with an α and β of 0.

DM must have some combination of other-regarding preferences and positive expectations that justify this choice.

When a DM chooses to cooperate in a one-shot PD game, she is choosing to accept the vulnerability of being defected upon and thus receiving the S(ucker) payoff. The motivation for doing so is provided by a sufficient belief that the other player will choose to cooperate (β), and that the marginal improvement ($R - S$) is enough to warrant this strategic "risk" over the strictly "safer" alternatives (T or P). This approach rationalizes trust based cooperative decisions. The emergent question then is: what combinations of prosocial preferences, positive expectations, and available payoffs, would induce a DM to trust and thus choose to cooperate in a social dilemma?

3. A model of trust based cooperation

We can make predictions about whether a DM will choose to cooperate or defect in a PD game by finding the maximum of the two strategies' expected utilities given the DM's preferences (α), beliefs (β), and the PD game's payoff structure (*TRPS*). The expected utility of choosing to cooperate in this context is:

$$u(C) = [\beta \cdot (R + \alpha \cdot R)] + [(1 - \beta) \cdot (S + \alpha \cdot T)]. \quad (2)$$

The expected utility of choosing to defect is:

$$u(D) = [\beta \cdot (T + \alpha \cdot S)] + [(1 - \beta) \cdot (P + \alpha \cdot P)]. \quad (3)$$

A DM will choose to cooperate when $u(C)$ is strictly greater than $u(D)$. Given this representation, we can determine the critical values of α and β that form a threshold between cooperation and defection as a subjective expected utility maximizing strategy.

4. Results

Fig. 2 shows a particular PD game with payoffs $T = 1, R = 0.8$, $P = 0.6$, and $S = 0$.The figure shows when one would expect a decision maker to cooperate in the specified PD given particular combinations of α and β values from the interval [0, 1] each. Coordinates on the Cartesian plane that are light indicate cases where $u(C) > u(D)$ and thus where cooperation is predicted. Conversely, coordinates shaded dark indicate cases where $u(C) < u(D)$ and where defection is predicted.

Given a perfectly selfish DM, we would not expect cooperation, no matter what the DM believes the other player will choose. However, if the DM is somewhat prosocial and has an alpha of 0.5 (this corresponds to an SVO angle[1] of about 27°), then we would expect the DM to cooperate only if he is at least 67% sure that the other player will also cooperate.

This *trust threshold* is clearly dependent on the particular PD game's payoff structure. A trust threshold (like that shown in Fig. 2) can be found with Eq. (4) given a PD game with payoffs *TRPS*, and a joint utility function like that in Eq. (1).

$$\beta_{crit} = \frac{P - S + \alpha P - \alpha T}{P + R - S - T + \alpha P + \alpha R - \alpha S - \alpha T}. \quad (4)$$

Next we consider the 26 PD games listed in Table 1. The model predictions for all of these different PD instantiations are displayed in Fig. 3. PD games that are characterized by the same K-index of cooperation are located in the same row. There are several things to note from this visual explanation. First, as the K-index increases, the size of the cooperate area (light shaded) generally increases as well. This makes intuitive sense as it indicates that those PD games that demand less SVO and/or less positive expectations in order to justify trust are also those where we would expect more cooperation to emerge. There is general agreement between K and the model about a particular PD's degree of *severity*. However the relationship between K, and the location and curvature of the *trust threshold* is not perfectly linear. The model differentiates between PD games that share the same K-index, and thus provides a more nuanced view of a PD game's degree of severity and may better predict when cooperation will emerge.

A novel index of cooperation can be derived from the model's predictions. The area of cooperation (CoopArea; i.e., the light area) is an additional indicator of a PD game's *severity*. Alternatively, one could also identify the value of α that is minimally sufficient to evoke cooperation given the principle of insufficient reason, i.e., given $\beta = 0.5$. Such an index (α_{crit}) can be understood as the answer to the question "how much SVO does a DM need in order to justify cooperation, given that the DM has no idea (e.g., a uniform prior) what the other player is going to choose?". This critical SVO level[2] can be found for a PD game with the following equation:

$$\alpha_{crit} = -\frac{P - R - S + T}{P - R + S - T}. \quad (5)$$

The rank correlations between the different PD games' α_{crit} values, the corresponding areas of cooperation, and the corresponding K-indexes are shown in Fig. 4, which also provides a visualization of the relation between these three indexes of cooperation as bivariate scatter plots.

Additionally, another well-known metric can also be viewed as an index of cooperation, namely the "Price of Anarchy" (PoA; see e.g., Mak & Rapoport, 2013). In the context of PD games, the PoA is simply the ratio $\frac{R}{P}$.Clearly, the indexes extracted from the

[1] α values can be translated into SVO angles by the following equation: $\tan(\alpha) = SVO^{\circ}$ and can be translated back the other way via $SVO^{\circ} = \arctan(\alpha)$. For values of α between 0 and 1, this is close to a linear transformation. For more information regarding computations of SVO angles, see Murphy, Ackermann, and Handgraaf (2011). Traditionally, SVO scores are reported as angles, but rescaling them produces values that are more readily interpretable.

[2] This critical value uses the joint utility function as stated in Eq. (1). A different utility function would yield a different formula for α_{crit}.

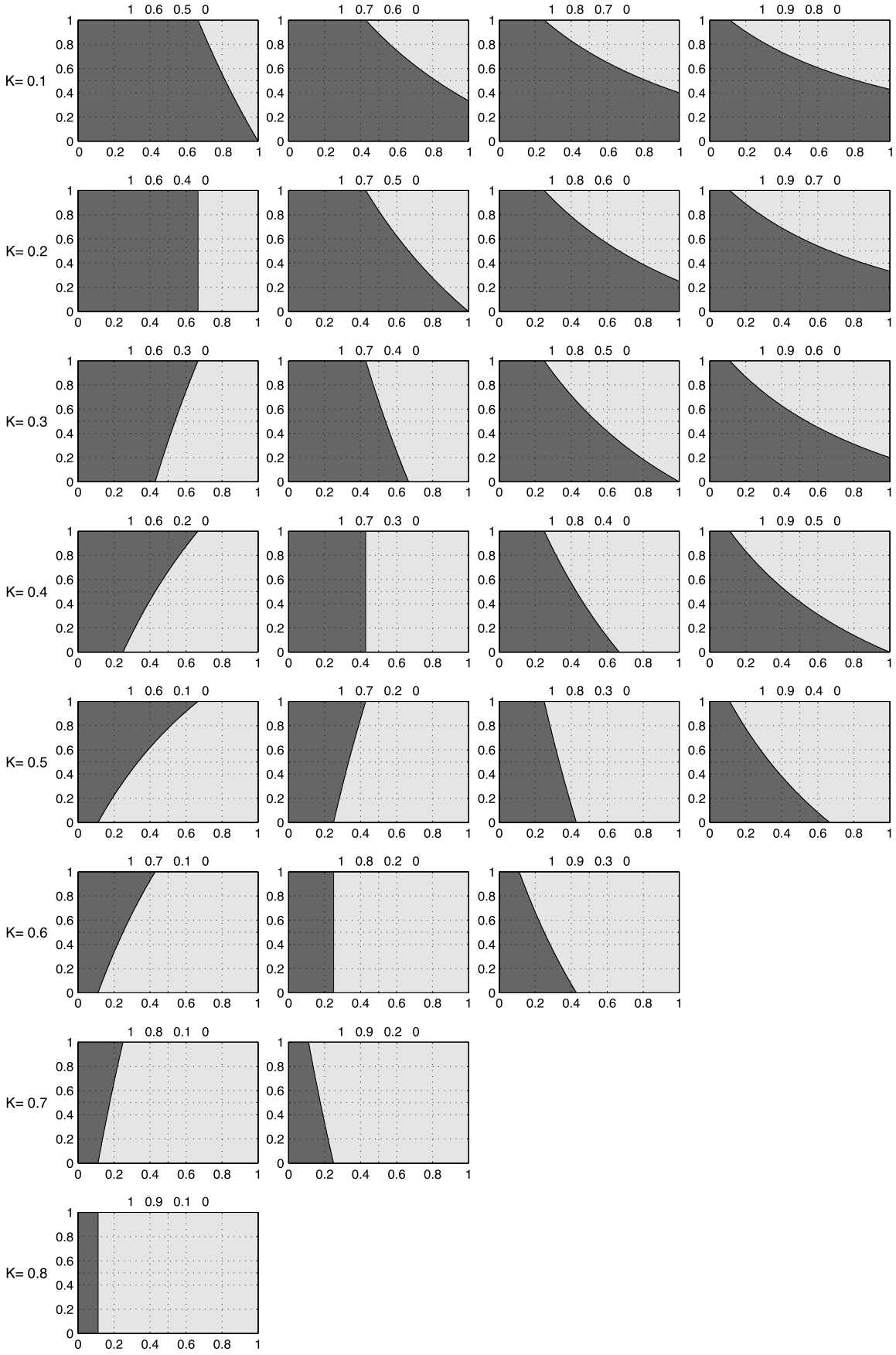

Fig. 3. Alpha–Beta plots showing trust thresholds for a variety of different PD games. Like Fig. 2, each subplot has an x-axis that corresponds to SVO (α) and a y-axis that corresponds to positive expectations (β). Rapoport's K-indexes are shown on the left and the payoffs (*TRPS*) that define each particular PD game are shown above each subplot. The light areas correspond to expected *cooperation*.

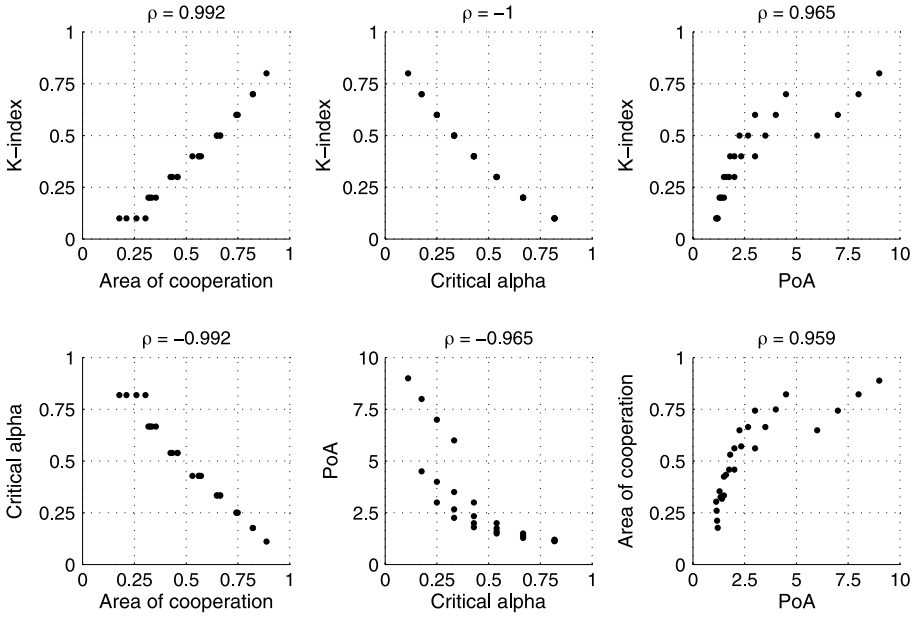

Fig. 4. The relations among the K-index, PoA, CoopArea, and α_{crit}. Each point is a particular PD game from Table 1 and the Spearman rank correlation coefficient is displayed at the top of each scatterplot.

model predictions are strongly related to the K-index of cooperation, which may provide some insight about why the K-index is predictive of cooperation rates in PD games in the first place, namely because it captures some of the psychology behind the interaction between preferences, beliefs, and payoffs. In fact, the K-index almost perfectly corresponds (negatively) to the minimum degree of prosociality required to trust and act cooperatively given the principle of insufficient reason. But the K-index is less sensitive than the CoopArea index because it only takes into account one of the interactions' three components (i.e., payoffs), while the area of cooperation is affected by three factors (i.e., preferences, beliefs, and payoffs).

A counterintuitive result of this model is that the slope of the trust threshold line is sometimes positive. This means that for some PD games the model predicts that DMs will be *more* likely to cooperate given *lower* expectations of the other player. This is a consequence of the simple joint utility model from Eq. (1). To ensure that only negatively sloping trust thresholds would emerge, one would need to implement a different social utility model. For example a joint utility function with a contingent component could be used (e.g., a DM's social preferences are positive if and only if the other player chooses C, zero otherwise). With the use of more complicated contingent social preference models, the shapes of the trust thresholds are transformed and have different properties, some of which may be desirable from a descriptive vantage.

5. Discussion

We have addressed several issues in the process of characterizing trust based cooperation in social dilemmas by drawing upon the standard 2×2 PD game as the most prominent exemplar. Our main findings can be summarized as follows.

First, trust based cooperation in a social dilemma, such as the PD game, is rationalizable within a subjective expected utility framework given that the DM has: (a) sufficiently prosocial preferences; (b) sufficiently positive expectations that the other player will choose to cooperate; and (c) the potential payoffs form a sufficiently non-severe game. This approach accommodates heterogeneity across decision makers in preferences and beliefs, and can constructively account for individual differences in observed choice behavior while still accommodating *homo economicus* as a literal corner case ($\alpha = \beta = 0$).

Second, "trust thresholds" can be derived in preference–belief (α, β) space that defines for which combinations of SVO and positive expectations, trust based cooperation is rationalizable and a best response. In the PD game, these thresholds are related to the K-index but are not identical. Further, the preference–belief space (Fig. 2) is a useful framework to consider individual differences, especially in cases where a DM is close to cooperating but not quite over the threshold. Effective welfare enhancing nudging would imply finding the minimum distance to "move" a DM across the trust threshold, and that requires interventions aimed at changing preferences or at changing beliefs. Knowing the shape of the trust threshold, and a DM's current location in preference–belief space, would be useful in designing efficacious nudges.

Third, the K-index can be regarded as the answer to the question, "what is the minimum level of SVO (i.e., other regarding preferences) a DM would require to justify cooperation given a uniform prior?" The K-index itself does not capture how more nuanced beliefs might induce a DM to choose to cooperate or defect, as very different PD games can have identical K-index values. Fig. 5 shows the full range of possible PD games and where the 26 games discussed previously are located in this normalized PD space. Implicit in this account is a particular joint utility function that is used to model a DM's social preferences, and the K-index happens to correspond to a simple model of other regarding preferences.

Fourth, we have developed normalized PD games by "feature-scaling" the payoffs. This kind of standardization is valuable for supporting the accumulation of knowledge by facilitating comparisons across different results from different studies. Independent research groups have used a wide variety of different PD games to date and due to differently scaled payoffs the similarity or disparity of these different games is obscured. Knowing where on the normalized PD map (Fig. 5) a certain game is located would be useful for disentangling the effects of payoffs from other mechanisms of interest.

Finally, our findings support cautioning against relying too much on theoretical predictions of behavior in strategic contexts from simulation studies that result from the investigation of a single PD game with *a particular* payoff structure. A simple decision model can make dramatically different predictions of

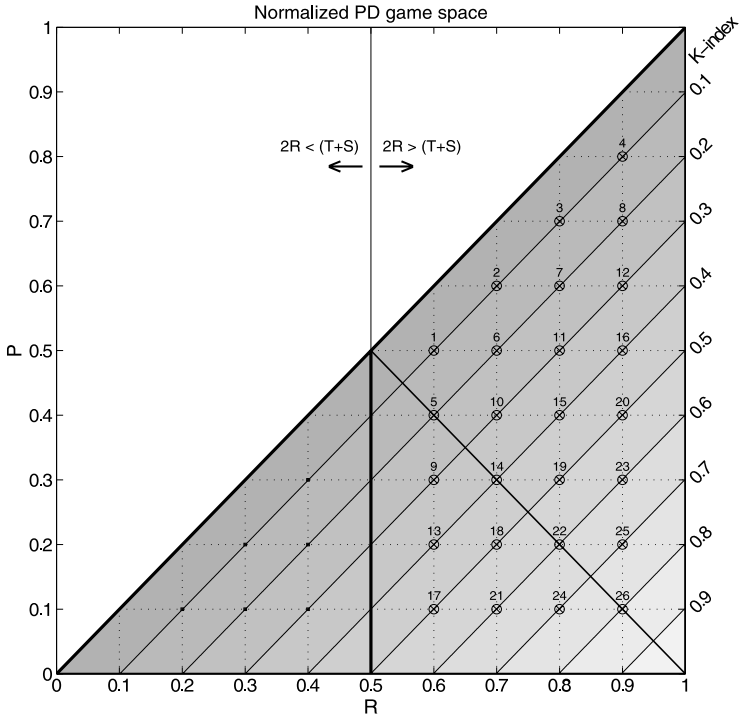

Fig. 5. The normalized space of PD games varying free parameters R and P, while fixing $T = 1$ and $S = 0$. The 26 PD games considered here (see Table 1) are marked with a $\otimes$ symbol. PD games above the -45° line from (1,0) to (0,1) have a trust threshold with a negative slope using the joint utility function from Eq. (1). PD games positioned directly on the -45° line (e.g., PD games 5, 14, 22, 26) have a perfectly vertical trust threshold, meaning that beliefs are not related to a DM's choice in those games. PD games below the -45° line have a positively sloped trust threshold.

behavior *within the same game* due to the interactions between the social utility model and the payoffs (see Fig. 3). Usually, researchers do not report comprehensive instantiations of a particular type of game and thus risk overgeneralizing results to other contexts when predictions may not even hold for different instantiations of a PD game with different payoffs. This pitfall may be particularly well hidden when choice behavior is modeled and predicted on the aggregate level, such that attention is drawn to the values of some parameters, the combination of which happens to describe aggregate behavior particularly well. However the goodness of fit may be a unique result and only occurs at the intersection of particular game payoffs and particular model parameters. Models that work well across the whole PD space would be more compelling and provide stronger evidence of model quality.

Acknowledgments

Thanks to Daniel Balliet and Robert ten Brincke for our discussions and valuable feedback on early versions of this paper.

References

Andreoni, J., & Miller, J. H. (2002). Giving according to GARP: An experimental test of the consistency of preferences for altruism. *Econometrica*, *70*(2), 737–753.

Au, W., & Kwong, J. (2004). Measurements and effects of social-value orientation in social dilemmas: A review. In R. Suleiman, D. Budescu, I. Fischer, & D. Messick (Eds.), *Contemporary psychological research on social dilemmas* (pp. 71–98). New York: Cambridge University Press.

Axelrod, R. (1967). Conflict of interest: An axiomatic approach. *Journal of Conflict Resolution*, *11*, 87–99.

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, *10*(1), 122–142.

Glöckner, A., & Hilbig, B. E. (2012). Risk is relative: Risk aversion yields cooperation rather than defection in cooperation-friendly environments. *Psychonomic Bulletin & Review*, *19*(3), 546–553.

Kreps, D. M., Milgrom, P., Roberts, J., & Wilson, R. (1982). Rational cooperation in the finitely repeated prisoners' dilemma. *Journal of Economic Theory*, *27*(2), 245–252.

Mak, V., & Rapoport, A. (2013). The price of anarchy in social dilemmas: Traditional research paradigms and new network applications. *Organizational Behavior and Human Decision Processes*, *120*, 142–153.

Murphy, R. O., & Ackermann, K. A. (2014). Social value orientation: Theoretical and measurement issues in the study of social preferences. *Personality and Social Psychology Review*, *18*(1), 13–41.

Murphy, R. O., Ackermann, K. A., & Handgraaf, M. J. J. (2011). Measuring social value orientation. *Judgment and Decision Making*, *6*(8), 771–781.

Rapoport, A. (1967). A note on the index of cooperation for Prisoner's Dilemma. *Journal of Conflict Resolution*, *11*(1), 100–103.

Rapoport, A., & Chammah, A. M. (1965). *Prisoner's dilemma: A study in conflict and cooperation*. Ann Arbor: University of Michigan Press.

Roth, A. E., & Murnighan, J. K. (1978). Equilibrium behavior and repeated play of the prisoner's dilemma. *Journal of Mathematical Psychology*, *17*, 189–198.

Rousseau, D., Sitkin, S., Burt, R., & Camerer, C. (1998). Not so different after all: A cross-discipline view of trust. *Academy of Management Review*, *23*(3), 393–404.

Steele, M., & Tedeschi, J. (1967). Matrix indices and strategy choices in mixed-motive games. *Journal of Conflict Resolution*, *11*(2), 198–205.

Van Lange, P. A. M. (1999). The pursuit of joint outcomes and equality in outcomes: An integrative model of social value orientation. *Journal of Personality and Social Psychology*, *77*(2), 337–349.

Vlaev, I., & Chater, N. (2006). Game relativity: How context influences strategic decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(1), 131–149.