University of Groningen

Explaining contributions to public goods

Dijkstra, Jacob

Published in:
Rationality and Society

DOI:
[10.1177/1043463111434702](https://doi.org/10.1177/1043463111434702)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2012

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

Dijkstra, J. (2012). Explaining contributions to public goods: Formalizing the social exchange heuristic. *Rationality and Society*, 24(3), 324-342. <https://doi.org/10.1177/1043463111434702>

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Explaining contributions to public goods: Formalizing the social exchange heuristic

Rationality and Society

24(3) 324–342

© The Author(s) 2012

Reprints and permission:

sagepub.co.uk/journalsPermissions.nav

DOI: 10.1177/1043463111434702

rss.sagepub.com

Jacob Dijkstra

University of Groningen/ICS, The Netherlands

Abstract

The public good game is a popular model of cooperation problems. Rational egoism predicts that in finitely repeated public good games no contributions are made. At least 4 observations are inconsistent with this prediction: contributions (i) are frequently positive, (ii) increase in the marginal rate of return of the public good, (iii) increase in the expected contributions of others, and (iv) decrease as the public good game is repeated. I build a rational choice model that explains these observations, extending and formalizing the social exchange heuristic. The model does not assume a specific utility function. I assume that total utility is a strictly increasing function of monetary and social utility, and then show that concave utility functions and a positive cross-partial derivate of total utility to monetary and social utility are sufficient conditions to arrive at a model compatible with the 4 observations. Directions for future (experimental) research are discussed.

Keywords

cooperation, public goods, social exchange heuristic, social preferences

Introduction

Public goods (PGs) constitute an important topic in the social sciences. Among other things, the PG framework is frequently used to model behavior in so-called

Corresponding author:

Jacob Dijkstra, Department of Sociology, University of Groningen/ICS, Grote Rozenstraat 31, 9712 TG Groningen, The Netherlands

Email: j.dijkstra@rug.nl

'tragedy of the commons' situations (e.g. Hardin, 1968, Milinski *et al.*, 2002). Sociologists draw attention to the link between 'social capital' and PGs. Thus, according to Putnam (1995, page 66, emphasis added) "'social capital" refers to features of social organization such as networks, norms, and social trust that facilitate *coordination and cooperation for mutual benefit.*" Coleman (1990, pages 310–311, emphasis added) argues that '[a] prescriptive norm that constitutes an especially important form of social capital within a collectivity is the norm that *one should forgo self-interest to act in the interest of the collectivity.*' Thus, according to these scholars, the ability of a community to efficiently produce PGs is an important element of its social capital.

The standard rational choice paradigm, based on rational egoism, predicts that in finitely repeated PGs no individual should contribute anything to the PG. However, there are at least 4 frequently made observations in one-shot or finite-shot PGs that are inconsistent with this prediction: (i) people frequently contribute positive amounts (e.g. Ledyard, 1995; Sally, 1995), (ii) people increase their contributions as the marginal rate of return of the PG increases (e.g. Camerer, 2003), (iii) people increase their contribution when they expect others will contribute more (e.g. Croson, 2007; Mulder *et al.*, 2006), and (iv) people decrease their contributions when the PG is played repeatedly a finite number of times with the same partners (e.g. Fehr and Gächter, 2002).

In the current paper I build a rational choice model based on the *social exchange heuristic* (SEH; Kiyonari *et al.*, 2000; Yamagishi *et al.*, 2007), that can explain these four observations. The SEH was conceived to explain cooperation in prisoner's dilemmas (PDs), and leads to the prediction that individuals will cooperate more when they expect others will cooperate. First I extend the theory to include PGs. Then I formalize the theory and show under which conditions the four observations mentioned above are compatible with the model.

Of course there is a host of social preference models (e.g. Bolton and Ockenfels, 2000; Fehr and Schmidt, 1999) that are applicable to the problem of contributions to public goods. However, as I expound below, there is particularly strong *direct* evidence for the individual preferences described by the SEH. Moreover, the SEH is buttressed by an evolutionary explanation, i.e., in terms of ultimate goals. It is therefore important that this simple but powerful theory be generalized to cover PGs, and formalized.

The paper that substantively comes closest to the current paper is Ambrus and Pathak (2011). It is instructive to briefly compare their approach to mine. Ambrus and Pathak (2011) model a finitely repeated public goods game, as I do. They set out to explain the fact that empirical contributions in such games are non-zero and declining over time. In order to do this they make a number of assumptions. They start by assuming there are two categories of players:

reciprocal and selfish players. They then make five additional assumptions about the 'reciprocity functions' that govern the behavior of the former category of players. These assumptions guarantee the existence of a unique subgame perfect Nash equilibrium with strictly positive contributions; a result that is arguably much stronger than the results of the current paper. This equilibrium exhibits decreasing contributions over time as desired, and is compatible with two additional frequently observed behavioral patterns: (i) lengthening the game increases aggregate contributions, and (ii) a 'surprise restart' of the game leads to a sudden increase in contributions (an effect known as 'the restart effect', see Andreoni (1988)). However, these results come at a price that clearly distinguishes their approach from mine.

The assumptions Ambrus and Pathak (2011) make on the reciprocity functions are very strong and to a large extent hard to justify substantively (the 'additive separability' and 'linearity' assumptions perhaps being the clearest examples of this). There is nothing wrong with this in principle, since the authors show how these assumptions lead to the desired results.¹ While this is certainly a very worthwhile endeavor and the paper of Ambrus and Pathak (2011) is a technical achievement, I take a different approach.

Below, I will make assumptions sparingly and I will try to argue that each assumption is empirically plausible by giving it substantive meaning or by citing experimental evidence that supports it. Perhaps most importantly, I will not assume that players in the model are necessarily in equilibrium. This will be explained in detail in the section describing the model.

There is one assumption that I will not try to justify. To render the whole modeling approach of the paper fruitful in the first place, I must assume twice differentiable (and therefore continuous) utility functions. All this will lead to results that are weaker than the existence of a unique subgame perfect equilibrium, but that are derived from assumptions that are easier to defend.²

In addition to the issue of making assumptions, I generally take a different approach from Ambrus and Pathak (2011). This paper formalizes the informal micro-theory of the SEH that has been offered to explain cooperation in social dilemmas. As I will argue below, the basic claim of this theory, that individuals prefer mutual cooperation over defecting on a cooperating partner, is empirically supported in neurological and experimental research. In the formalization I assume that all agents' utility functions have both traditional monetary arguments as well as social arguments. The SEH impacts on the latter. Given these basic theoretical choices, I try to find the *minimal empirically defensible* assumptions on agents' utilities that can accommodate the 4 basic facts about repeated public goods games mentioned above.

The contribution of the current paper is thus to show how empirical facts can be explained with a model based on an established micro-theory, making

Table 1. A prisoner’s dilemma game (PDG) and an assurance game (AG); adapted from Kiyonari *et al.* (2000).

		PDG		AG	
		Player A		Player B	
		Cooperate	Defect	Cooperate	Defect
Player A	Cooperate	2, 2	0, 3	2, 2	0, 1
Player B	Defect	3, 0	1, 1	1, 0	1, 1

Note: First entry in each cell indicates payoff of player B; second entry indicates payoff of player A.

minimal assumptions. Of course, this model will not be the definitive model on this topic. However, it is by constructing and comparing different models, based on different theoretical premises that we can hope to advance our understanding of how real humans make decisions in social dilemmas.

In the next section I briefly present the theoretical reasoning underlying the SEH and extend it to PGs. In the subsequent section I formalize these ideas and derive my results. The paper is concluded with a brief discussion.

Theory

The central claim of the SEH is that people generally have a cognitive bias in the information processing of social exchange, in the sense that they tend to perceive a prisoner’s dilemma game (PD) as if it were an assurance game (AG).

Consider the PD of Table 1. It can be regarded as a discrete, two-player version of a PG. Thus, according to the standard rational egoist model and assuming players A and B make a one-time simultaneous decision to choose either Cooperate or Defect, each is guided by her self-interest to choose Defect, since this is her best choice regardless of the choice of the other player. This individually rational course of action yields them both a payoff of 1. Had they both chosen Cooperate, however, they each would have gotten the larger payoff of 2. Thus, whereas both choosing Cooperate leads to the social optimum, self-interest nonetheless leads players to choose Defect. This makes the PD a true *dilemma*: the social optimum is at odds with individual rationality. Note how in the PD the expectation of the behavior of the other player plays no role in the determination of an individual’s rational action. It is best to play Defect regardless of what the other chooses, so expecting that the other will play Cooperate should have no effect on one’s choice.

Kiyonari *et al.* (2000), however, claim that many people actually perceive this PD as an AG, also depicted in Table 1. Note how in the AG a player will want to play Cooperate provided the other player also plays Cooperate. If the other chooses Defect, however, a player will also prefer to play Defect. Thus, contrary to what was the case with the PD, in the AG the expectations concerning the behavior of the other *do* play an important role in determining an individual's rational course of action. If she expects the other to play Cooperate, an individual should also play Cooperate; if she does not expect the other player to play Cooperate, she should play Defect. Note that the SEH does not imply that when encountering a PD players are mistaken regarding the material payoffs in the matrix. The SEH contends that all things considered players prefer the situation in which both players get the 2 payoff as a result of successful mutual cooperation over the situation in which they themselves get the 3 payoff while their partner gets the 0 payoff, as a result of the fact that they have cheated on the cooperating partner.

Kiyonari *et al.* (2000) base their claim on an evolutionary argument which is itself based on theory and findings of Cosmides and Tooby (1992). The latter authors argue that through evolution humans have been equipped with a cognitive module for social exchange that includes the ability to detect cheaters. The argument for the existence of this cheater detection ability is fortified by direct experimental results (i.e., the existence of cheater detection ability is not just inferred from the fact that humans successfully cooperate in social dilemmas; see, for instance, Cosmides and Tooby (1992, chapter 3)). The SEH argument is that assuming humans have cheater detection ability, cheaters will not be accepted as exchange partners (even by fellow cheaters, who would rather exchange with cooperators). Therefore, cheater detection ability is only useful for individuals who can be accepted as an exchange partner, i.e., individuals (non-cheaters) who want to cooperate with others. Thus, if humans have developed cheater detection ability, they must also have developed the intention to cooperate. Hence the cognitive bias of evaluating the outcome of mutual cooperation higher than the outcome that results when one defects on a cooperating partner.³

The SEH can be perceived as a heuristic implementation of the famous tit-for-tat (TFT) strategy (Axelrod, 1984). This strategy also 'wants' to cooperate with a cooperating partner but 'wants' to defect on a defecting partner. Yamagishi *et al.* (2007) argue that the SEH is adaptive because there is typically a lot of uncertainty in social exchange relations regarding whether the encounter is one-shot or repeated. Incorrectly inferring that the relation is one-shot and defecting is then very risky, as it might spoil a potentially beneficial long-term social relationship. In addition, one's defective behavior might become known to other individuals, tarnishing one's reputation.

Especially in the close-knit human groups of the ancestral environment, this is a much larger risk than the one entailed by incorrectly inferring that the relation is long-term and cooperating.

Kiyonari *et al.* (2000) cite evidence showing that many subjects in PD experiments indeed rate the desirability of the outcome of mutual cooperation higher than that of the outcome resulting from defecting on a cooperating partner. Similarly, Rilling *et al.* (2002) offer questionnaire and neurological evidence from an experiment using the PD of Table 1 (with payoffs being dollar amounts), showing that mutual cooperation is evaluated as the most desirable outcome by participants, and is associated with consistent activation in brain areas related to reward processing. Finally, the SEH is also consistent with the account of Frank (1988) concerning the need for a solution to the 'impulse control problem'. Humans excessively discount future payoffs compared to current ones. This makes enduring cooperation hard to sustain, since myopic individuals are not very susceptible to the threat of future losses due to a breakdown of cooperation. The mental transformation of a PD into an AG causes individuals to actually *prefer* mutual cooperation in the present, thus dissipating the impulse control problem.

The SEH was conceived to deal with two-player PDs, and can be captured by saying that a player with SEH preferences tries to avoid both being a 'free-rider' and being a 'sucker'. The first problem one encounters when extending the SEH from PDs to PGs is the question what defines a sucker and a free-rider when contribution decisions are continuous instead of discrete. A second problem is that we will want to apply the theory and model to general N -player PGs, and not just to two-player games.

To solve these problems I assume that players have some well-defined expectations concerning the contributions of all others. These expectations may be the result of some social norm, of introspection (e.g. Ellers and Van der Pool, 2010), of previous experience or of still other factors. Based on these expectations, I assume that each player defines some 'appropriate level of contribution' for herself. This appropriate level will in general be influenced by some social norm, but I am not assuming that all players will hold the same appropriate level of contribution when they have the same expectations.

Based on the PD evidence cited above I expect that the appropriate level of contribution will be strictly positive when the player expects others to make strictly positive contributions. Subsequently, I model the SEH by assuming that the appropriate level of contribution serves as a benchmark for the player to distinguish between free-rider behavior and sucker behavior. Contributing more than the appropriate level would make the player feel a sucker, whereas contributing less than the appropriate level would make the player feel a free-rider. With these very general assumptions in hand I now turn to the modeling.

Model

The rational individual that I model below has well-defined expectations about the contribution decisions of others, and based on these expectations predicts the utility she will experience for each level of her own contribution. When dealing with repeated play, at the end of this section, I will assume that a player's expectations of the contributions of others are updated after each round of play. Importantly, I am assuming that players have definite expectations about the contributions of others, and I am not modeling these expectations as random variables with probability distributions.

A linear public good

Assume there is a linear one-shot PG, to which the focal player can contribute part of her initial endowment e . Let the predicted monetary outcome of this decision for the focal player be $M = M(x, \bar{y}) = (e - x) + \frac{cx}{N} + \sum_j y_j \frac{c}{N}$, where

x is the focal player's contribution, $\bar{y}$ is the vector of predicted contributions of the other players, j sums over the other players, N is the number of players, and c is the *efficiency factor* of the public good, with $1 > c > 0$. This yields

$$\frac{\partial M}{\partial x} = \frac{c}{N} - 1 < 0 \text{ and } \frac{\partial M}{\partial y_i} = \frac{c}{N} > 0, \text{ for any } y_i \text{ that is a component of } \bar{y}.$$

Furthermore, linearity means that second order (cross-)partial derivatives are

$$\text{zero: } \frac{\partial^2 M}{\partial x^2} = \frac{\partial^2 M}{\partial y_i^2} = \frac{\partial^2 M}{\partial y_i \partial x} = 0.$$

Monetary utility

I assume that the player's utility consists of two components, namely monetary utility and social utility which are represented by functions that are assumed continuous and (at least) twice differentiable. With the social utility function I will model the SEH preferences. Write the monetary utility as $u_M = u_M(M)$,

and assume that the player prefers more money to less: $\frac{du_M}{dM} > 0$. Then we

can derive the standard rational egoist result of no contribution in a one-shot PG by computing the first derivative of u_M to x : $\frac{du_M}{dM} \frac{\delta M}{\delta x} < 0$. Thus any

marginal increase in contribution decreases utility, and the only rational thing to do is to contribute nothing. Note that, in accordance with the PD in the previous section, the expectations concerning the contributions of others ($\bar{y}$) have no effect on the contribution decision.

Positive contributions

To formalize the SEH for the PG case, I assume that the player determines some appropriate level of contribution, based on the predicted contributions of others: $E = E(\bar{y})$. I assume that contributions *below* E are defined by the focal player as free riding. Contributing *more* than E would make the focal player feel like a ‘sucker’. Concerning E , I assume that $\frac{\partial E}{\partial y_i} \geq 0$ for any y_i that

is a component of $\bar{y}$. Thus, the appropriate level of contribution does not decrease as the expected contribution by some other player increases.

Now write the focal player’s social utility as $u_S = u_S(x, E(\bar{y}))$. I formalize the SEH preferences as $\frac{\partial u_S}{\partial x} > 0$, if $x < E(\bar{y})$, $\frac{\partial u_S}{\partial x} < 0$, if $x > E(\bar{y})$, and $\frac{\partial u_S}{\partial x} = 0$ if $x = E(\bar{y})$. Thus, when contributing less than the appropriate level, the player feels like a free-rider and can increase her social utility by increasing her contribution. When contributing more than the appropriate level, the player feels like a sucker and can increase her social utility by decreasing her contribution. These assumptions are in line with for instance Kurzban *et al.* (2001) who argue that in public good games ‘players want to contribute at the same level as others in their group, preferring neither to free ride nor to be free ridden.’

Now write total utility as a strictly increasing, continuous and (at least) twice differentiable function of both monetary and social utility: $U = U(u_M, u_S)$. The assumptions on the utility functions imply that U is either decreasing in x , or first increasing and then (beyond a certain point) decreasing in x . In both cases U is quasiconcave in x and a Nash equilibrium is guaranteed to exist (see for instance Theorem 1.2 in Fudenberg and Tirole, 1991, page 34). However, I do not contend that players necessarily play an equilibrium.

An equilibrium implies that each player correctly forecasts the contributions of all other players. In a game with a dominant strategy (such as the public goods game with only monetary utility, or the ‘untransformed’ PD) this is not problematic, since a player’s best reply does not depend on what she expects others will do. However, in a model such as the current, where a player’s best reply might depend on her expectations about others, the assumption of equilibrium play becomes very restrictive. It requires that players be flawless prophets. Note in this respect that I am trying to build a model that can accommodate actual experimental findings. Therefore, I just assume that players have definite expectations about contributions of others, not that these expectations will be met. In the words of the famous Thomas theorem: ‘If men define situations as real, they are real in their consequences.’

To find the optimal level of x (given the focal player's expectations), take the derivative of U to x and set it to 0 (first order condition of optimality):

$$F = \frac{\partial U}{\partial u_M} \cdot \frac{du_M}{dM} \cdot \frac{\partial M}{\partial x} + \frac{\partial U}{\partial u_S} \cdot \frac{\partial u_S}{\partial x} = 0 \quad (1)$$

It follows from (1) and my previous assumptions that nonzero contributions are *feasible*: the first term of (1) is always negative, but if $\frac{\partial U}{\partial u_S} \cdot \frac{\partial u_S}{\partial x}$ is positive and sufficiently large, it will be optimal to invest some nonzero amount in the PG. Whether in any empirical case the focal actor actually makes nonzero contributions depends on the relative strength of the monetary and social utility components, as can be seen from (1). The focal player's expectations give rise to an appropriate level of contribution E . Then E and $U = U(u_M, u_S)$ must be such that at least for levels of contribution close enough to 0, the social part of (1) is positive *and* larger in absolute value than the monetary part of (1). This means that the focal actor will contribute a strictly positive amount. As I said above, in a Nash equilibrium the focal actor's expectations about the contributions of others are correct, and the amount she contributes is exactly equal to the amount other actors expected from her.

By the SEH preferences we must have that $x < E(\bar{y})$, for $\frac{\partial U}{\partial u_S} \cdot \frac{\partial u_S}{\partial x}$ to be positive. Thus it follows from (1) and the SEH preferences that the player will always 'underinvest' compared to E . Thus, if players with SEH preferences care just a little for money (i.e., the monetary part of (1) is nonzero), they will 'free-ride' a little.

The above implies that if (1) holds, $\frac{\partial u_S}{\partial x} > 0$: the player can increase her *social* utility (but not her *total* utility) by contributing more to the PG. In line with SEH preferences, I assume that in this case $\frac{\partial u_S}{\partial E} < 0$: given contribution x , social utility will decrease as the appropriate level of contribution increases.

The marginal rate of return of the PG

In the remainder of this paper I consider an actor for whom, given her (possibly erroneous) expectations, there is an interior solution to the utility maximization problem: the player contributes a strictly positive amount that is smaller than her entire endowment e , and (1) is satisfied.

To investigate the effect of changes in the efficiency factor c on x , let (1) define x as an implicit function of c . Then by the implicit function theorem

$$\frac{dx}{dc} = - \frac{\frac{\partial F}{\partial c}}{\frac{\partial F}{\partial x}}. \text{ A necessary condition for this derivative to exist is that } \frac{\partial F}{\partial x} \neq 0.$$

It turns out that this is also a sufficient condition for x to locally be a function of c . For (1) we find that:

$$\begin{aligned} \frac{\partial F}{\partial x} = \frac{\partial^2 U}{\partial u_M^2} \cdot \left(\frac{du_M}{dM} \cdot \frac{\partial M}{\partial x} \right)^2 + 2 \frac{\partial^2 U}{\partial u_S \partial u_M} \cdot \frac{\partial u_S}{\partial x} \cdot \frac{du_M}{dM} \cdot \frac{\partial M}{\partial x} \\ + \frac{\partial^2 U}{\partial u_S^2} \cdot \left(\frac{\partial u_S}{\partial x} \right)^2 + \frac{\partial U}{\partial u_M} \cdot \frac{d^2 u_M}{dM^2} \cdot \left(\frac{\partial M}{\partial x} \right)^2 + \frac{\partial U}{\partial u_S} \cdot \frac{\partial^2 u_S}{\partial x^2} \end{aligned} \quad (2)$$

In principle all the elements of the right-hand side of (2) can separately be either positive, zero or negative, implying that without making further assumptions the sign of $\frac{\partial F}{\partial x}$ depends on the relative magnitude of these elements.

Since we want to draw general conclusions concerning the relationships between contribution x and the other parameters and variables in the model, *without having to work with specific functional forms*, we will have to make some minimal assumptions on the utility functions. It turns out that we need no more than two such assumptions to guarantee a definite sign of (2).

The first assumption is that $\frac{\partial^2 U}{\partial u_S \partial u_M} \geq 0$. Three widely used classes of utility functions, namely linear functions, Cobb-Douglas functions and constant elasticity of substitution (CES) functions, meet this requirement.⁴ Substantively, the assumption means there is a trade-off between the monetary and the social utility component. If the *level* of one of these components increases, the *marginal impact* of the other component on the actor's well-being also increases. For instance, as the amount of money the actor earns increases she becomes more concerned with meeting the appropriate level of contribution.

The second assumption is that marginal utility is decreasing. Thus, both earning more money and making a contribution closer to the appropriate level increases the actor's well-being, but as the level of these utility components rises, the impact of additional increases on well-being becomes smaller. That this assumption is necessary, can be seen as follows.

For linear utility functions (i.e., constant marginal utility) all second order (partial and cross-partial) derivatives are zero, which will make (2) zero. Thus,

for linear utilities we cannot generally define x as a local function of c , and we cannot draw general conclusions regarding the effect of changes in c on x . Of course, for linear utilities there will not be an interior solution to (1), in the first place. For increasing marginal utility, second-order derivatives will be positive. This implies that the second element of (2) will be negative and the other elements will be positive. Thus, the sign of (2) depends on the *actual* utility functions chosen, and we again cannot draw *general* conclusions.

For decreasing marginal utility, which is a standard microeconomic assumption for modeling individual decisions, we can draw general conclusions about the effects of changes in several variables on x . In this case all second order derivatives are negative, and (2) becomes negative. In the remainder of the paper I thus make these two assumptions, unless stated otherwise.

Concerning the effect of the efficiency factor c on F , I assume that E and $\bar{y}$ are not explicit functions of c . Thus, neither the appropriate level of contribution nor social utility itself are directly affected by the efficiency of the PG. Given (1) we then find that

$$\frac{\delta F}{\delta c} = \left[\frac{\delta^2 U}{\delta u_M^2} \left(\frac{du_M}{dM} \right)^2 + \frac{\delta U}{\delta u_M} \frac{d^2 u_M}{dM^2} \right] \frac{\delta M}{\delta x} + \frac{\delta U}{\delta u_M} \frac{du_M}{dM} \frac{1}{N} > 0.$$

This yields $\frac{dx}{dc} = - \frac{\frac{\delta F}{\delta c}}{\frac{\delta F}{\delta x}} > 0$. Thus we have established that given an interior

solution to (1), players with decreasing marginal utility and SEH preferences

for whom $\frac{\partial^2 U}{\partial u_S \partial u_M} > 0$ (e.g. Cobb-Douglas or CES functions) will increase

their contribution if the efficiency factor of the public good increases.

While increasing the efficiency factor c enhances the PG's marginal rate of return, increasing the number of players decreases it. To find the effect of changes in the number of players on the contribution decisions, let (1) implicitly

define x as a function of N . By the implicit function theorem $\frac{dx}{dN} = - \frac{\frac{\delta F}{\delta N}}{\frac{\delta F}{\delta x}}$.

In an analysis similar to the previous one we find that given (1), $\frac{\delta F}{\delta N} < 0$.

Thus, $\frac{dx}{dN} = - \frac{\frac{\delta F}{\delta N}}{\frac{\delta F}{\delta x}} < 0$, and we have established that given an interior solu-

tion to (1), players with decreasing marginal utility and SEH preferences for

whom $\frac{\partial^2 U}{\partial u_S \partial u_M} > 0$ (e.g. Cobb-Douglas or CES functions) will decrease

their contribution if the number of players increases.

The two results from this subsection show that a model based on SEH preferences is compatible with the generally observed pattern of increasing contributions in the marginal rate of return of PGs.

Contributions of others

To investigate how a change in the expected contributions of others affects the contribution of the focal player, assume that (1) implicitly defines x as a function of y_j , for any y_j that is a component of $\bar{y}$. Once more by the

implicit function theorem we have $\frac{dx}{dy_j} = -\frac{\frac{\partial F}{\partial y_j}}{\frac{\partial F}{\partial x}}$. Given (1) we find:

$$\begin{aligned} \frac{\partial F}{\partial y_j} = & \frac{\partial^2 U}{\partial u_M^2} \cdot \left(\frac{du_M}{dM}\right)^2 \cdot \frac{\partial M}{\partial x} \cdot \frac{\partial M}{\partial y_i} + \frac{\partial^2 U}{\partial u_S \partial u_M} \cdot \frac{\partial u_S}{\partial E} \cdot \frac{\partial E}{\partial y_i} \cdot \frac{du_M}{dM} \cdot \frac{\partial M}{\partial x} \\ & + \frac{\partial U}{\partial u_M} \cdot \frac{\partial^2 u_M}{\partial M^2} \cdot \frac{\partial M}{\partial x} \cdot \frac{\partial M}{\partial y_i} + \frac{\partial^2 U}{\partial u_M \partial u_S} \cdot \frac{\partial u_S}{\partial x} \cdot \frac{du_M}{dM} \cdot \frac{\partial M}{\partial y_i} \\ & + \frac{\partial^2 U}{\partial u_S^2} \cdot \frac{\partial u_S}{\partial x} \cdot \frac{\partial u_S}{\partial E} \cdot \frac{\partial E}{\partial y_i} + \frac{\partial U}{\partial u_S} \cdot \frac{\partial^2 u_S}{\partial E \partial x} \cdot \frac{\partial E}{\partial y_i} \end{aligned} \quad (3)$$

For players with decreasing marginal utility and $\frac{\partial^2 U}{\partial u_S \partial u_M} > 0$, the first 5 terms of (3) are positive. Moreover, since decreasing marginal utility means concave utility functions, $\frac{\partial^2 u_S}{\partial E \partial x}$ will be positive too: an increase in E will

increase the distance between E and x , and by concavity this will increase the first derivate of u_S to x . Thus, we find that (3) is positive, and

$$\frac{dx}{dy_j} = -\frac{\frac{\partial F}{\partial y_j}}{\frac{\partial F}{\partial x}} > 0. \text{ This establishes that given an interior solution to (1),}$$

players with decreasing marginal utility and SEH preferences for whom

$$\frac{\partial^2 U}{\partial u_S \partial u_M} > 0 \text{ (e.g. Cobb-Douglas or CES functions) will increase their con-}$$

tribution if they expect other players to increase their contributions.

Declining contributions over time in finitely repeated play

If the public good game would be infinitely repeated, we could in principle find Nash equilibria with strictly positive contributions that do not decline over time, depending on the actual utility functions (about which we made no assumptions). As discussed in the introduction, Ambrus and Pathak (2011) show that given sufficient assumptions we can guarantee the existence of a unique subgame perfect equilibrium in a finitely repeated game. However, as I argued above for the one-shot case, equilibrium play assumes players correctly predict the contributions of the other players. In the finitely repeated game this means players correctly forecast (or somehow know) the entire trajectory of contributions of all players in all repetitions. Thus, in the finitely repeated game equilibrium play imposes an even stronger assumption on players' predictive abilities than in the one-shot game. Since I am trying to understand observed behavior in experiments, I therefore do not assume equilibrium play. Given that players in the SEH model have expectations about the contributions of others to which they respond, this leads me to take a 'belief learning' approach in which players update their expectations based on their experiences in previous rounds (see Camerer, 2003, chapter 6, for an overview of learning models in game theory).

To study the process of adjustment of x when the public good game is repeated a finite number of times, I have to make two assumptions. The first concerns the updating of expectations. The simplest such assumption is that the expected contributions by others are equal to the most recently observed contributions by others. In addition, I have to make an assumption on the function E . A straightforward assumption is that E equals the average of the expected contributions of others. Substantively, these two assumptions mean that players (i) expect others to behave in the current round as they behaved in the previous round, and (ii) feel that contributing *more* than the average makes one a 'sucker', whereas contributing *less* than the average constitutes 'free-riding'.

I start with the fact that according to the SEH preferences, for each player k $x_k < E(\bar{y})$. Now let i be a player with the highest contribution in period t , $x_{i,t}$. Then, by the assumptions concerning the updating of expectations

and the observation that given SEH preferences we have $x_k < E(\bar{y})$,

$$x_{i,t} \geq \frac{\sum_{j \neq i} x_{j,t}}{N-1} > x_{i,t+1}. \text{ Moreover, for all other players } j \neq i \text{ we have}$$

$$x_{i,t} \geq \frac{\sum_{k \neq j} x_{k,t}}{N-1} > x_{j,t+1}. \text{ Thus, for all periods } t, \text{ the maximum contribution in}$$

period $t + 1$ is strictly lower than the maximum contribution in period t . In fact, it is straightforward to see that since the appropriate contribution level in the current period equals the average contribution from the previous period, the average contribution itself decreases over time.

First of all, note that this result holds regardless of whether or not $\frac{\partial^2 U}{\partial u_S \partial u_M} > 0$, and regardless of whether or not players have decreasing marginal utilities. Thus, these two assumptions that I had to make in the one-shot case to draw conclusions about the effects of changes in the efficiency factor or the expected contributions of others, can be dropped now. Secondly, the conclusion that the maximum contribution in period $t + 1$ is strictly lower than the maximum contribution in period t also follows for other specifications of E . In particular, if E is either the maximum, the median, or the minimum of all contributions from the previous round, the results still hold. Thus, even if players feel that the appropriate level of contribution in round t is equal to the maximum contribution made in round $t - 1$, the maximum contribution still steadily declines over time. This declining pattern is driven by my two initial modeling choices: (i) social utility is derived from comparison of own contribution to the appropriate contribution level, with deviations in either direction lowering social utility, and (ii) total utility is composed of a monetary and a social component between which there is a trade-off.

One must be careful with the conclusions drawn from this result. All it establishes is that the highest contribution (and the average, median or minimum, depending on the choice of E) will steadily decline over time and for instance not that it will eventually be zero. Moreover, these results are derived only for special assumptions on the updating process and on the forming of the appropriate level of contribution. Thus, the simple dynamics described in this section are based on a basic learning assumption: players update their beliefs by setting them equal to observed behavior from the previous round and subsequently play a best reply given their new beliefs (a process also known as Cournot dynamics). However, as before no specific assumptions were made concerning the functional forms of players' utilities.

Discussion

In this paper I developed a rational choice model based on the social exchange heuristic. The model is compatible with frequently observed behavior of individuals in public good games, such as contributing non-zero amounts even in one-shot situations. These behaviors cannot be explained by the traditional rational egoist assumptions on individual preferences.

In the model development I extended the social exchange heuristic, originally conceptualized for explaining cooperation in the prisoner's dilemma, to the public good game. To do this, I introduced players' expectations concerning the contributions of others and assumed these expectations to be the basis for an 'appropriate level of contribution'. This raises the interesting questions of whether players actually hold such an appropriate contribution level, and what factors affect it. I believe these questions should and can very well be investigated experimentally.

Marwell and Ames (1981) already found that many of their experimental participants held definite conceptions of what constituted a 'fair' level of contribution to a public good project. Excluding economics graduates and participants for whom it was individually rational to contribute, they found that 'more than three out of four thought that 'about half' or more of a person's resources should be contributed, and more than one out of four thought people who were fair would contribute *all* of their tokens' (Marwell and Ames, 1981, page 308). Moreover, Marwell and Ames (1981) also show that holding a high appropriate level of contribution is not sufficient in itself to induce high contributions. In particular, they find that participants who contribute the most were those who both defined high levels of contribution as fair *and* indicated that they were concerned with fairness when making their contribution decision. This suggests that the appropriate level of contribution is defined relatively independently of the weight that it carries in the contribution decision. This is consonant with the model from this paper, that defines the appropriate level of contribution separately from the weight of social utility in determining total utility (i.e., the first derivative of total utility to social utility).

A basic claim of the current model is that individuals contribute less than their appropriate level of contribution (unless they do not care for monetary utility at all, in which case their contribution is equal to their appropriate level). Whether this is actually true is a question that should be experimentally investigated. Note that a truly heuristic interpretation of the SEH hypothesis would predict that contributions will be equal to their appropriate levels: the social exchange features of the environment are cues that trigger the heuristic and individuals act accordingly. However, in the current model I assumed

that the heuristic affects *preferences*, which are then included in a rational choice decision. Thus, testing whether contributions indeed fall short of appropriate levels sheds light on whether the contribution decision is truly heuristic or more of a dual nature.

In any case, as noted in the introduction, contribution decisions *are* affected by the public good's efficiency factor, which suggests that either the heuristic affects preferences (as in the current model) or the payoffs of the game are part of the cues that determine whether or not the heuristic is activated. These cues would then be interpreted differently by different individuals. Indeed, Simpson (2004) found that a given experimental PD game was interpreted by only part of participants as an AG, while the rest played it as a genuine PD.

A strong point of the model developed in this paper is that it does not assume any specific functional form of players' utilities. It turns out that assuming decreasing marginal utility and a positive cross-partial second order derivative of total utility to monetary and social utility are sufficient conditions to allow definite conclusions in the current model. This raises the interesting questions of the extent to which these assumptions hold for real individuals and how behavior is affected by changes in these assumptions. These question can be addressed either theoretically, by making more specific assumptions on the utility functions, or by conducting public good experiments that include measurements of participants' utilities, or impose certain utility functions in an incentive compatible way.

The model of this paper is of course limited in a number of respects. First of all, the model assumes that the expectations of the contributions of others are 'deterministic'. Thus, an interesting theoretical extension of the current model would be to allow the expectations concerning other players' contributions to be probabilistic. This would allow the derivation of hypotheses concerning the effects of distributional properties of the expectations on for instance the appropriate level of contributions, and the weight of social utility in the contribution decision. For example, one would expect that as an individual is less certain about her expectations of the contributions of others, the associated appropriate level of contribution is assigned less weight in her own contribution decision. It would be very interesting to test hypotheses like these experimentally.

Another limitation of the current paper is that the analysis of the repeated game is admittedly very rudimentary. However, the goal of the present paper was not to introduce an elaborate learning model, but to show how with relatively few assumptions, non-standard preferences can explain observed behavior. In the repeated game I had to make additional assumptions on the learning dynamics and on the appropriate level of contribution. However,

the assumptions of decreasing marginal utility and a positive cross-partial second order derivative of total utility to monetary and social utility could be dropped. The non-standard preferences I assumed were based on the social exchange heuristic that is supported by important direct evidence, cited in the paper. Moreover, as indicated above, the current model offers starting points for the investigation of factors that are theoretically expected to be important for behavior in public good games. These factors include expectations of others' contributions, and the existence and determinants of the 'appropriate' level of contribution or norm. Note that in an orthodox rational egoist model, none of these factors is assumed to matter in finitely repeated public good games.

Notes

- 1 One could argue that the fourth assumption of Ambrus and Pathak (2011) (that reciprocal players reciprocate contributions of selfish players that are known to be motivated by selfishness), is self-contradictory, since reciprocity is triggered by a 'kind and costly' act of another agent. The same is true of their final equilibrium result: it is an equilibrium in which reciprocal players knowingly reciprocate selfish players' instrumental behavior. This is of course a consequence of their assuming the existence of two distinct categories of players.
- 2 In fact, one of the assumptions made by Ambrus and Pathak (2011), that reciprocal player do not 'overreciprocate', follows from the model in the current paper.
- 3 Concerning the truly *heuristic* nature of the SEH, I don't assume a dual mode decision model of either rational choice or heuristic, but rather a mix: the heuristic implies certain preferences, and given these preferences a rational decision is made. In his respect the model developed in this paper might be closer to the 'preferences over intentions' approach of for instance Rabin (1993) and Falk *et al.* (2008).
- 4 Linear utilities have the form $u = \sum_i a_i x_i$, Cobb-Douglas utilities have the form

$$u = k \prod_i x_i^{b_i}, \text{ and CES utility functions have the form } u = k \left(\sum_i c_i x_i^{-a} \right)^{-\frac{b}{a}}, \text{ where}$$

the x s denote amounts of different commodities. In economic production theory the positivity of the cross-partials of these functions (then conceived as production functions) is sometimes referred to as Wicksell's Law.

References

- Andreoni J (1988) Why free ride? Strategies and learning in public goods experiments. *Journal of Public Economics* 37: 291–304.
- Ambrus A and Pathak PA (2011) Cooperation over finite horizons: A theory and experiments. *Journal of Public Economics* 95: 500–512.

- Axelrod R (1984) *The Evolution of Co-operation*. New York, NY: Basic Books.
- Bolton GE and Ockenfels A (2000) ERC: A theory of equity, reciprocity, and competition. *The American Economic Review* 90(1): 166–193.
- Camerer CF (2003) *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton, NJ: Princeton University Press.
- Coleman JS (1990) *Foundations of Social Theory*. Cambridge, MA: The Belknap Press of Harvard University Press.
- Cosmides L and Tooby J (1992) Cognitive adaptations for social exchange. In: Barkow JH, Cosmides L, and Tooby J. (eds.) *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*, pp. 163–228, New York: Oxford University Press.
- Croson RTA (2007) Theories of commitment, altruism and reciprocity: Evidence from linear public good games. *Economic Inquiry* 45(2): 199–216.
- Ellers J and Van der Pool NCE (2010) Altruistic behavior and cooperation: The role of intrinsic expectation when reputational information is incomplete. *Evolutionary Psychology* 8(1): 37–48.
- Falk A, Fehr E, and Fischbacher U (2008) Testing theories of fairness: Intentions matter. *Games and Economic Behavior* 62: 287–303.
- Fehr E and Gächter S (2002) Altruistic punishment in humans. *Nature* 415: 137–140.
- Fehr E and Schmidt K (1999) A theory of fairness, competition and cooperation. *The Quarterly Journal of Economics* August: 817–868.
- Frank RH (1988) *Passions within Reason: The Strategic Role of the Emotions*. New York: W.W. Norton & Company.
- Fudenberg D and Tirole J (1991) *Game Theory*. Cambridge, MA: MIT Press.
- Hardin, G (1968) The tragedy of the commons. *Science* 162: 1243–1248.
- Kiyonari T, Tanida S, and Yamagishi T (2000) Social exchange and reciprocity: Confusion or a heuristic? *Evolution and Human Behavior* 21: 411–427.
- Kurzban R, McCabe K, Smith VL, and Wilson BJ (2001) Incremental commitment and reciprocity in a real-time public goods game. *Personality and Social Psychology Bulletin* 27(12): 1662–1673.
- Ledyard JO (1995) Public goods: A survey of experimental research. In: Kagel JH and Roth AE (eds.) *The Handbook of Experimental Economics*, pp. 111–194. Princeton, NJ: Princeton University Press.
- Marwell G and Ames RE (1981) Economists free ride, does anyone else? *Journal of Public Economics* 15: 295–310.
- Milinski M, Semmann D, and Krambeck HJ (2002) Reputation helps solve the ‘tragedy of the commons’. *Nature* 415: 424–426.
- Mulder L, Van Dijk E, De Cremer D, and Wilke HAM (2006) Undermining trust and cooperation: The paradox of sanctioning systems in social dilemmas. *Journal of Experimental Social Psychology* 42: 147–162.
- Putnam RD (1995) Bowling alone: America’s declining social capital. *Journal of Democracy* 6(1): 65–78.
- Rabin M (1993) Incorporating fairness into game theory and economics. *American Economic Review* 83: 1281–1302.

- Rilling JK., Gutman DA, Zeh TR, Pagnoni G, Berns GS, and Kilts CD (2002) A Neural Basis for Social Cooperation. *Neuron* 35: 395–405.
- Sally D (1995) Conversation and cooperation in social dilemmas: A meta-analysis of experiments from 1958 to 1992. *Rationality and Society* 7(1): 58–92.
- Simpson B (2004) Social values, subjective transformations, and cooperation in social dilemmas. *Social Psychology Quarterly* 67(4): 385–395.
- Yamagishi T, Terai S, Kiyonari T, Mifune N, and Kanazawa S (2007) The social exchange heuristic: Managing errors in social exchange. *Rationality and Society* 19(3): 259–291.