Weak moral motivation leads to the decline of voluntary contributions

Marc Willinger

Weak moral motivation leads to the decline of voluntary contributions

August 2, 2009

Abstract

We develop a model that accounts for the decay of the average contribution observed in experiments on voluntary contributions to a public good. The novel idea is that people's moral motivation is "weak". Their judgment about the right contribution depends on observed contributions by group members and on an intrinsic "moral ideal". We show that the assumption of weakly morally motivated agents lead to the decline of the average contribution over time. The model is compatible with persistence of over-contributions, variability of contributions (across and within individuals), and the "restart effect". Furthermore, it offers a rationale for conditional cooperation.

Keywords: Conditional cooperation, voluntary contributions, moral motivation, experiments on public goods games.

JEL: H00, H41, C72.

1 Introduction

Several experimental studies have documented strong empirical regularities in linear public goods experiments including (1) the fact that people contribute more than predicted by the standard theoretical model; and (2) that average contribution declines steadily over time when the game is repeated under a finite horizon.[1] In contrast to the huge amount of literature concerned with the puzzling fact that subjects over-contribute with respect to their Nash contribution, the attempts made to explain the decline of the average contribution observed in most linear public goods experiments are not clear-cut. Understanding the driving forces behind this decay has obvious policy implications. For instance, if the decay is initiated by a few free-riders, policies oriented towards punishing individuals that free-ride on contributors could be contemplated. However, if most people are weakly motivated to contribute, policies aiming at fostering individuals' motivation, through public information campaigns, are recommended. The attempts at explanation boil down to three major explanations : learning, strategic play and other regarding preferences including reciprocity (conditional cooperation).

According to the learning hypothesis over-contributions in early rounds arise because subjects are confused and make errors. As time elapses, and feedback from past rounds becomes available, they realize that they could earn more by over-contributing less, and adjust their current contribution accordingly. However, available evidence about learning suggests that it plays a limited role in the decay. Neugebauer et al. (2009a) found that repetition without feedback has no effect on average contribution which seems to suggest that there is no learning by introspection. Decay arises only when information about the contribution of group members is provided. In contrast, Houser & Kurzban found a sharper decay when a subject plays against a computer program.[2] Nevertheless, the learning hypothesis seems incompatible with the "restart effect" found by Andreoni

[1] See Ledyard (1995) for a review of this literature published prior to 1995; See also Andreoni, 1995; Croson, 1996; Gaechter & Fehr, 1999; Keser & van Winden, 2000; Fehr & Gaechter, 2000; Masclet et al., 2003; Carpenter, 2007; Sefton, Shupp & Walker, 2007; Herman, Toeni & Gaechter, 2008.

[2] In their computerized treatment human subjects were told that the 3 other members of their group

(1988).[3][4]

The second candidate explanation for the decay in average contribution is reputation. The hypothesis of "strategic play" is based on the idea that players take into account future interactions when choosing their current contribution. Therefore, in early periods they have an incentive for establishing a cooperative reputation, by making a large contribution. The justification of the strategic hypothesis is based on the "crazy player" assumption (Kreps et al.,1982), or equivalently on the lack of common knowledge of rationality. If (rational) players believe that there is a crazy player in the group who contributes positively in period 1, it becomes rational for them to play a trigger strategy in early periods and to mix over the strategy space as the repeated game approaches the final period.[5] Andreoni (1988) offered the first test of the reputation hypothesis by comparing the av-

where computerized players, which followed a predetermined contribution path. The latter was chosen to be equal to $\frac{3}{4}$ of the average contribution observed in the human condition, and the computers' current choice was announced at the beginning of each round. Overcontribution did not vanish even after 10 rounds. Houser & Kurzban's conclusion is that more than 50% of the overcontribution is due do confusion.

[3]After having played 10 rounds of a linear public good game, Andreoni told his subjects that a new sequence of 10 rounds will be played. Surprisingly, he observed that the average contribution of the first restart period does not differ from the average first period contribution in the initial sequence.

[4]Anderson et al. (2004) developed a more sophisticated theory of learning based on quantal response. Anderson et al. (2004) show that the dynamic process of individual contributions follows the well-known Fokker-Planck equation and converges to the logit equilibrium distribution of contributions. Despite its mathematical elegance, the model has two limitations : first, the model does not explain why contributions are sensitive to the remaining number of periods as observed in partner sessions (due to the neglect of strategic interactions across periods), and second, it assumes that subjects are able to best-respond to the stochastic distribution of other players' contributions by choosing a stochastic distribution over their strategy space, which requires a high degree of sophistication for each player to form expectations about other players' choice probabilities. While the model nicely describes the patterns observed in the data, "as if" subjects' behaviour was of the QRE type, it is not very plausible from a behavioural point of view.

[5]Strategic play is compatible with the fact that most subjects over-contribute in early periods and switch to their Nash contribution at some later period (see e.g. Isaac et al. (1994), Laury (1997), Keser & vanWinden (2000)). In early periods they signal a desire to cooperate, but as the end of the game approaches their incentive to do so vanishes, and they switch to their Nash contribution. The evidence that the decay is slower in longer games (Isaac et al. 1994) is compatible with the strategic hypothesis.

erage contributions of partner groups with stranger groups. Since there is no incentive to develop a cooperative reputation among strangers, one should observe higher over-contributions in partner-groups than in stranger-groups, especially in early periods of the repeated game. Surprisingly, Andreoni (1988) found that strangers contribute more than partners, that the difference in average contribution increases over time, and that complete free-riding is significantly more frequent in the partner treatment. These findings seem to undermine the reputation hypothesis as a plausible explanation of the decay in average contributions.[6]

The third explanation is rooted in other-regarding preferences. Altruism [7] can account for the fact that people contribute more than their pure self-interested contribution, but cannot explain the decay in contribution. A more plausible explanation is conditional co-operation, i.e. the fact that people choose to cooperate, depending on previously observed decisions of others or on beliefs about their decisions (e.g. Keser and van Winden, 2000; Fischbacher, Gaechter and Fehr, 2001; Croson, 2007; Fischbacher & Gaechter, 2009). Up to recently, reciprocity theories tended to attribute the decay to preference heterogeneity : reciprocal cooperative players are mixed with selfish agents who free-ride on others' contributions. In a given period, a reciprocal player who either observes that his contribution is above the average, or who expects others to contribute less, will reduce his contribution, and therefore the mean contribution declines. But heterogeneity per se is

[6]Note however that there is no clear evidence that stranger contribute significantly more than partners. Indeed, several studies have also found the opposite, showing that partners contribute significantly more than strangers (see for example Croson, 1996, Fehr & Gaechter, 2000; Masclet et al., 2003). However these studies also indicate that average contribution under stranger matching condition still decline over time as the game is repeated, which cannot be explained by the reputation hypothesis.

[7]One potential explanation relies on the idea that people cooperate because they "take pleasure in others' pleasure" (see e.g. Dawes and Thaler,1988). Theory of altruism presented by Andreoni and Miller (1996) assumes that an altruistic player's utility increases not only in his own payoff but also in the other players' payoffs. Two forms of altruism are generally given in the literature: "pure altruism" (individuals care about others' payoffs) and "warm-glow" altruism (individuals enjoy contributing per se). However, both pure and impure altruistic motives cannot adequately describe the decline of contribution observed in public goods experiments. Indeed why would altruistic motives vanish over time?

neither a necessary nor a sufficient condition for the decline.[8] Heterogeneity can either reinforce or attenuate the tendency for the decline, but is not the central driving force of the process. According to Fischbacher & Gaechter (2009) imperfect conditional cooperation is the main driving force behind the decay : "Many people's desire to contribute less than others, rather than changing beliefs of what others will contribute over time". There is strong experimental evidence for such imperfect reciprocity (Fischbacher & Gaechter, 2009, Fischbacher et al., 2001) or selfishly-biased reciprocity (Neugebauer et al., 2009). A conclusion from this short literature review is that imperfect reciprocity is the main driving forces underlying the decay, while learning and strategic behavior reinforce this tendency.

In this paper, we propose a new model of behavior, compatible with the imperfect conditional cooperation hypothesis, that accounts for the decline of average contribution. Precisely, it is based on the idea that agents set their moral target by relying on two dimensions : a "morally ideal contribution" (see Brekke et al., 2003, Nyborg, 2000) and the observed contributions of others. The assumption that people rely on a morally ideal contribution is defended by commitment theories (see e.g. Croson 2007). Based on Kantian reasoning, these theories assume that individuals make "unconditional" commitments to contribute to the public good. Our originality is to assume that for most people such commitments are weak in the sense that they are sensitive to the observation of others' actions.

[8]It is not necessary because if reciprocators contribute a little less than the observed (or expected) average contribution, decay can arise even in a population composed exclusively of non-selfish agents. A population of identical non-selfish agents, but slightly selfishly oriented, is enough to provoke the decline.

To see why it is not sufficient, let us define a perfect reciprocator as a player who matches the average group contribution of the previous period. Consider a case involving only two players : a perfectly reciprocal player and an unconditional player who contributes a fixed amount in each period. As the game is repeated the contribution of the reciprocal player converges to the fixed contribution of the unconditional player, with a slope that depends on the initial contribution of the reciprocator. Hence the average contribution could increase! The example can be easily extended to any mixed population of any finite size composed of perfect reciprocators and unconditional players. Adding noisy players who contribute a random amount does not prevent that the mean contribution either decays or increases.

In a contribution context, individuals might therefore be tempted to revise their preferred contribution after observing others' contributions. The extent of such a revision typically varies across individuals : strongly morally motivated agents will closely stick to their ideal contribution, while weakly motivated agents are prone to revise their morally ideal contribution whenever they observe a gap between their own and others' contributions. Our idea of weakly morally motivated agents captures a large spectrum of contribution behaviors : at one extreme, unconditional contributors, who always stick to their ideal contribution whatsoever, and at the other extreme pure reciprocators who always match the observed group contribution. Free-riding behavior is a particular case of unconditional behavior : always contributing zero in a linear public good game whatever the other group members do. Our idea, is that most people are of the "mixed" type, i.e. their actual contribution is the outcome of a deliberative process through which their preferred contribution is balanced against others' observed average contribution. The assumption of weakly morally motivated agents offers a possible justification for imperfect reciprocal behavior that we call *action-based reciprocity* : individuals' ideal contribution is sensitive to others' observed contributions. Alternatively, individuals might also decide about their contributions by relying on their expectations about others' contributions (*belief-based reciprocity*). However, such beliefs are themselves revised according to observed contributions.

The model is primarily designed to explain the decline of average contributions in linear public goods experiments. However, it is also compatible with several other stylized facts observed in experiments on voluntary contributions to a public good. For linear public goods experiments these facts may be summarized as follows : a) subjects contribute half of their endowment in the first period, b) the average contribution tends to decline as the game is repeated, c) there is significant over-contribution in the final period, d) there is high variance of individual contributions, e) most subjects adjust their contribution from one period to the next. Similar observations have been made for interior Nash equilibria and for interior dominant strategy equilibria, except that stylized fact a) should be understood as "subjects contribute mid way between the equilibrium contribution and the socially optimum contribution" (see Sefton & Steinberg, 1996, Laury & Holt, 1998).

Although we are mainly concerned with facts b) and e), our model is also compatible with facts a), c) and d).

Section 2 introduces the concept of weak moral motivation and shows its implications for the dynamics of average contributions in a simple linear public good model with myopic agents. Section 3 extends the results to non-myopic agents. Our assumptions and results are discussed in section 5 and contrasted with other models, with a particular attention to Kandori (2002), Klumpp (2005) and Ambrus & Pathak (2007) that are closely linked to our idea. Section 6 concludes.

2 Weak moral motivation and voluntary contributions

Consider n agents, indexed $i = 1, ..., n$, who can contribute voluntarily to a public good. Each of them has an endowment w_i, which he can split between his contribution to the public good, x_i, and the consumption of private goods, $w_i - x_i$. Using the notation $x_{-i} = \sum_{j \neq i} x_j$, the cardinal representation of agents' preferences with moral motivation is :

$$U^i(x_i, x_{-i}, \widehat{x}_i) = w_i - x_i + \beta_i(x_i + x_{-i}) - v_i(x_i - \widehat{x}_i) \ . \ \ i = 1, ..., n, \tag{1}$$

where $\beta_i \in]0, 1[$ is the marginal utility from consuming $G = x_i + x_{-i}$, the public good[9].

Agent i's moral motivation is embodied in the function $v_i(.)$, where $\widehat{x}_i$ stands for his moral obligation. Her loss of utility attached to any deviation from her moral obligation is $v_i(x_i - \widehat{x}_i)$. This function is assumed to be convex. In addition two natural assumptions

[9] The results of the present paper also hold when preferences are captured by quadratic utility functions:

$$U^i(x_i, x_{-i}, \widehat{x}_i) = \alpha_i(w - x_i) - (w_i - x_i)^2 + \beta_i(x_i + x_{-i}) - \frac{v_i}{2}(x_i - \widehat{x}_i)^2.$$

This family of functions, which allows for a dominant strategy equilibrium with strictly positive contributions, has been documented in the experimental literature by relatively few papers (Keser, 1996, Bracht et alii, 2008).

about $v_i(.)$ are as follows:

Assumption 1 $v_i(0) = 0,\ v_i(x_i - \widehat{x}_i) > 0$ *iff* $x_i \neq \widehat{x}_i$.

Assumption 2 $v_i'(.) \gtreqless 0 \ \Leftrightarrow\ x_i - \widehat{x}_i \gtreqless 0$

The first assumption is obvious: a departure from one's moral obligation entails a loss of utility. The second assumption means that, starting from a situation where agent i contributes less (more) than her moral obligation, a marginal increase of x_i reduces (increases) her loss of utility.

We shall conceptualize the *weak moral motivation* of each agent as a combination of two logics : an autonomous logic and the logic of social influence. The autonomous logic is captured by an ideal, or "ethical", level of contribution noted $x_i^* \geq 0$. For instance, it could correspond to a Pareto optimal level of contributions, i.e. contributing w_i for each i. Such autonomous logic can be grounded on a Kantian Categorical Imperative, or on an unconditional commitment to a contribution (Laffont 1975, Harsanyi 1980). The second logic, our originality, captures social influences *via* the average contribution observed in the immediate past, $\overline{x}_{t-1} \geq 0$. The group contribution is publicly observed after each period. Each player can therefore compare his contribution to the average group contribution. Discovering that her own contribution differs from the group contribution eventually leads her to judge that the society is less (or more) deserving than she initially though, and consequently to revise her moral obligation $\widehat{x}_i$. For example a player who discovers that he has contributed more (less) than the average might decrease (increase) his current moral obligation. This "weakness" in moral motivations echoes popular and well-documented ideas among moral philosophers and moral psychologists. Those scholars will see some bearings with the *internalist* versus *externalist moral motivation* debate that discusses the interactions between individual motivations and moral judgments (see for instance Zangwill, 2003). Do moral ideals motivate people's decisions or are moral opinions an ex-post rationalization largely dictated by social influences and circumstances? Our concept

is also in keeping with their discussion about the *weakness of will* (which they refer to as *akrasia*, see Holton 2007), an over-readiness to abandon or revise one's moral resolution. Do people change their mind in the course of interactions with others or is their will insufficiently strong to overcome the cost of self-enforcing their moral judgement? The standard decision theory used in economics typically flattens those aspects. The goal of the so-called *behavioral economics* program is to bring such aspects into economics. We will get back briefly to those subtleties in the conclusion of the paper.

Overall, the qualified moral obligation, $\widehat{x}_{it}$, is defined as a function of the aforementioned variables:

$$\widehat{x}_{it} = \begin{cases} x_i^*, & t = 0, \\ M^i\left(x_i^*, \overline{x}_{t-1}\right), & t = 1, 2, ... \end{cases}$$

where function $M^i(.,.)$ is discontinuous at $t = 0$, for there is no previous observations at that date that could be used to qualify the autonomous ethical level.

The moral obligation function satisfies the intuitive properties:

Assumption 3 $\frac{\partial M^i}{\partial x_i^*} = M_1^i \geq 0,\ \frac{\partial M^i}{\partial \overline{x}_{t-1}} = M_2^i \geq 0,$

Assumption 4 $x_{it}^{\min} = \min\left\{x_i^*, \overline{x}_{t-1}\right\} \leq M^i\left(x_i^*, \overline{x}_{t-1}\right) \leq x_{it}^{\max} = \max\left\{x_i^*, \overline{x}_{t-1}\right\}.$

Also, it is assumed that the aggregate qualified moral obligation is bounded above by the aggregate autonomous moral obligation:

Assumption 5 $\sum_i M^i\left(x_i^*, a\right) \leq \sum_i x_i^* = G^*\ , \quad \forall a \geq 0.$

Assumption 5 plays a key role for the dynamics of contributions in the linear public good game. According to this assumption aggregate contributions cannot exceed the aggregate initial moral motivation. For instance, if $x_i^* = w_i$ for all i, actual contributions are necessarily bounded by aggregate endowment. In a more general sense, Assumption 5 means that players' moral motivation is not grounded on utopia but on realism and

feasibility. However, while Assumption 5 holds in aggregate, it needs not be true at individual levels, *i.e.* $M^i(x_i^*, a) > x_i^*$ for some i and some a is a possibility. Assumption 5 is further discussed in Section 3.

Example 1 *An illustration of a weak moral motivation function is the following:*

$$\begin{aligned} \widehat{x}_{it} &= (1-\theta_i)\, x_i^* + \theta_i \overline{x}_{t-1} \ , \qquad \theta_i \in [0,1] \ , \\ &= x_i^* - \theta_i \left(x_i^* - \overline{x}_{t-1}\right) \ . \end{aligned}$$

The weight $1-\theta_i$ *may be interpreted as the "strength" of agent i's moral motivation. If* $\theta_i = 0$ *agent i has as strong moral motivation : he never deviates from his ideal contribution, whatever the observed average contribution by other members of his group. On the other hand, an agent for whom* θ_i *is close to 1 will strongly revise her initial moral ideal, whenever her current contribution differs from the average group contribution. Assuming the above revision rule, a purely reciprocal player can be defined as a player for whom* $\theta_i = 1$*, while a unconditional free-rider is defined by* $\theta_i = 0$ *and* $x_i^* = 0$.

If the contribution game is played only once, player i has a dominant strategy to contribute less than his moral motivation. He chooses x_i to solve :

$$\max_{x_i} \ w_i - x_i + \beta_i \left(x_i + x_{-i}\right) - v_i \left(x_i - x_i^*\right).$$

The first order condition gives :

$$-1 + \beta_i = v_i' \left(x_i - x_i^*\right)$$

At equilibrium the agent equalizes the marginal material cost of a contribution $(\beta_i - 1)$ to the marginal moral cost of a deviation from her moral ideal $(v_i'(x_i - x_i^*))$, which implies that her actual contribution is less than her moral ideal :

$$x_i = x_i^* + \left(v_i'\right)^{-1} (\beta_i - 1) < x_i^* \ ,$$

since $\beta_i < 1$, $(v_i')^{-1}(\beta_i - 1) < 0$.

While our definition of weak moral motivation is related to a player's sensitivity to social influence, the above result shows that there is also a "private component" that drives agents' decisions. The latter can be thought as the temptation to deviate from the moral ideal, in order to increase one's material utility. Indeed the individual chooses her optimum level of contribution by equalizing the marginal material cost of a contribution to the marginal moral cost of deviating from the moral ideal. As we shall see below, this selfish bias towards the material payoff plays also a role in the decay of average contributions.

3 Repeated play with myopic contributors

Assume now that the contribution game is played a finite number of periods. We assume that in each period players rely on their current updated moral motivation, which is determined by the observed average contribution of the previous period. A key assumption for this section, which is relaxed later in the paper, is that players do not take into account their influence on other players' future moral motivation when choosing their contribution. We define therefore, for each period of time, a Myopic Nash Equilibrium (MNE)[10] as a profile of contributions such that each agent's contribution maximizes his own current utility, given the other agents' contributions:

$$\max_{x_{it}} \; w_i - x_{it} + \beta_i (x_{it} + x_{-it}) - v_i (x_{it} - \widehat{x}_{it}) .$$

From the first order conditions, interior decisions solve:

$$-1 + \beta_i = v_i' (x_{it} - \widehat{x}_{it}) \;\; , \qquad \forall i, \forall t,$$

thus individual equilibrium contributions at period t are:

$$x_{it} = M^i (x_i^*, \overline{x}_{t-1}) + (v_i')^{-1} (\beta_i - 1) \;\; , \quad \forall i \; .$$

[10] This concept is not ours. In particular it has been used extensively in the literature on processes (see Drèze and De la Vallée Poussin, 1977, for instance).

The revision rule for the moral motivation naturally leads to an interpretation of *action-based* reciprocity: current period contributions are partly determined by past observed contributions (and partly by individuals' moral motivation). The revision rule can be easily adapted to capture *belief-based* reciprocity. If we substitute "others' average contribution in the previous period" by the "expectation about others' contributions". The revision rule becomes $\widehat{x}_{it} = M^i\left(x_i^*, \overline{x}_{it}^e\right)$,where $\overline{x}_{it}^e$ is agent i's expectation of the average contribution of other players for period t. The interpretation is now that agent i determines her current contribution by taking into account her expectation about the (current) contribution of other players and her initial moral motivation. The two formulations do not fundamentally differ if we assume that individuals' expectations are positively related to their observed contributions of other group members, *i.e.* if $\overline{x}_{it}^e = f(\overline{x}_{it-1})$, with $f'(.) > 0$. With this assumption we restrict our interpretation to action-based reciprocity.

Proposition 1 *At a MNE, the level of public good is non increasing over time. If Assumption 5 is verified with a strict inequality, then the level of public good is strictly decreasing over time.*

Proof. The proof is established recursively. Note first that $G_1 = \sum_i x_i^* + \sum_i \left(v_i'\right)^{-1}(\beta_i - 1) \leq \sum_i x_i^*$ because $\sum_i \left(v_i'\right)^{-1}(\beta_i - 1) \leq 0$. Then observe that $G_2 = \sum_i M^i\left(x_i^*, \frac{G_1}{n}\right) + \sum_i \left(v_i'\right)^{-1}(\beta_i - 1) \leq G_1$, because $\sum_i M^i\left(x_i^*, \frac{G_1}{n}\right) \leq \sum_i x_i^*$ by Assumption A_5.

Assume the property $G_t \leq G_{t-1}$ holds for $t = 3, ..., k$, for some k. To complete the proof, it must be established that $G_{k+1} \leq G_k$ This is straightforward, for if the property is true until $t \leq k$, it follows that $G_{k+1} = \sum_i M^i\left(x_i^*, \frac{G_k}{n}\right) + \sum_i \left(v_i'\right)^{-1}(\beta_i - 1)$ is not larger than $G_k = \sum_i M^i\left(x_i^*, \frac{G_{k-1}}{n}\right) + \sum_i \left(v_i'\right)^{-1}(\beta_i - 1)$, because each $M^i(.,.)$ is an increasing function of its second argument and this argument has fallen, $G_k \leq G_{k-1}$.

To obtain the second claim of the proposition, repeat the same logic using strict instead of large inequalities. ■

Assumption 5 turns out to be crucial in explaining the decay of aggregate contributions. It can receive two justifications. First, the assumption is necessarily satisfied in period 1,

since the selfish bias curbs downwards each individual's contribution with respect to her initial ideal contribution. Of course, this fact does not preclude that some players adjust their moral ideal upwards. Assumption 5 can therefore be thought in the following way : downwards adjustments by high-motivated agents always loom larger than upwards adjustment by low-motivated agents. Second, and more generally, it is reasonable to set as an upper limit to the aggregate moral ideal. Budget constraints are on obvious reason. But more importantly, the fact that group interactions occurs only over a finite number of periods, sets a natural upper boundary on individual revised moral ideals, whenever these are influenced by social interactions.

We now turn to another important regularity in public good experiments, namely that in the last period of the finitely repeated game, subjects over-contribute significantly compared to the Nash prediction (see Holt & Laury, 2008)[11]. The dynamics of aggregate equilibrium contributions $G_t = \sum_i x_{it}$ are:

$$G_t = \sum_i M^i \left(x_i^*, \frac{G_{t-1}}{n} \right) + \sum_i \left(v_i' \right)^{-1} \left(\beta_i - 1 \right). \tag{2}$$

The dynamic process in (2) can eventually reach a level of contribution equal to zero, the free-riding equilibrium in standard linear public good's games. To account for over-contributions in our framework, two additional assumptions on the moral motivation function are required : the first one stipulates that an increase of the previous level of public good has a less than proportional positive effect on the levels of weak moral motivations:

Assumption 6 $M_2^i \leq 1$.

The second requires the moral motivation to be strong enough to induce a positive level of public good at a MNE even if the previous observable level was zero, i.e. even with an extremely adverse social influence.

[11] The authors summarize the experimental finding about end game effects as follows : "As the total number of rounds in a session is varied, contributions towards the end of the session are no lower in long time horizon experiments (40 to 60 rounds) than in short time horizon experiments (10 to 20 rounds").

Assumption 7 $\sum_i M^i(x_i^*, 0) \geq -\sum_i (v_i')^{-1}(\beta_i - 1)$.

Theorem 1 *Under Assumptions 5, 6 and 7, the sequence of public good levels converges to a unique positive interior level* $G^\infty \in]0; \sum_i x_i^*[$.

Proof. Under Assumption 6, the right hand side of the dynamics (2) is a contraction. Therefore, according to Banach's fixed point theorem : *i*) the dynamics (2) has a unique steady state, *ii*) the sequence converges towards this steady state. Assumptions 5 and 6 respectively discard the zero and full contributions corner stationary points. ■

Proposition 2 *Under Assumption 6, the higher the autonomous ethical level* x_i^**, the higher the long run level of public good* G^∞.

Proof. The long run level of public good solves

$$G^\infty = \sum_i M^i\left(x_i^*, \frac{G^\infty}{I}\right) + \sum_i (v_i')^{-1}(\beta_i - 1). \tag{3}$$

Using the implicit function theorem:

$$\frac{dG^\infty}{dx_i^*} = \frac{M_1^i}{1 - \frac{\sum_h M_2^h}{I}} > 0$$

under Assumption 6. ■

Proposition 3 *Under Assumption 6, the higher the marginal utility of the public good* β_i*, the higher the long run level of public good* G^∞.

Proof. Using (3) and the implicit function theorem again, one finds:

$$\frac{dG^\infty}{d\beta_i} = \frac{1}{\left(1 - \frac{\sum_h M_2^h}{I}\right) v_i''} > 0,$$

since in the denominator the first term between bracket is positive under Assumption 6 and the second term is also positive, for $v_i(.)$ is a convex function. ■

4 Forward-looking contributors

In this section we relax the assumption of myopic behavior in order to investigate how forward looking behavior affects the dynamics of average contribution. Forward looking players take into account the impact of their current contribution on other players' revised moral motivation in future periods, and are aware that other players try to influence their own future moral motivation. Such mutual influence, might eventually lead to an increase in average contributions. We show however that, under reasonable assumptions, the average contribution still declines over time.

Let T be the number periods during which agents interacts and $0 < \sigma \leq 1$ their common discount factor. We define players' intertemporal utility as :

$$\sum_{t=0}^{T} \sigma^t U^i \left(x_{it}, x_{-it}, \widehat{x}_{it} \right) \ , \ i = 1, ..., n,$$

where

- $U^i \left(x_{it}, x_{-it}, \widehat{x}_{it} \right) = w_i - x_{it} + \beta \left(x_{it} + x_{-it} \right) - \frac{v_i}{2} \left(x_{it} - \widehat{x}_{it} \right)^2$, v_i a positive scalar,
- $\widehat{x}_{it} = \left(1 - \theta_i \right) x_i^* + \theta_i \overline{x}_{t-1} \ = M^i \left(x_i^*, \overline{x}_{t-1} \right), \quad \theta_i \in [0, 1]$.

We will consider a Markov perfect equilibrium (MPE) for this dynamic public good game. Reasoning backward, in the last period agent i takes as given x_{-iT} and solves:

$$\max_{x_{iT}} \ w_i - x_{iT} + \beta_i \left(x_{iT} + x_{-iT} \right) - \frac{v_i}{2} \left(x_{iT} - M^i \left(x_i^*, \overline{x}_{T-1} \right) \right)^2 .$$

He has a dominant strategy, configured by the average contribution inherited from the previous period:

$$x_{iT} = \frac{-\left(1 - \beta \right)}{v_i} + M^i \left(x_i^*, \overline{x}_{T-1} \right) \ \equiv g_{iT} \left(\overline{x}_{T-1} \right).$$

It is worth noting that

$$g'_{iT} = M_2^i \left(., . \right) = \theta_i. \tag{4}$$

Those equilibrium strategies can be plugged back into the last period utility, giving each agent's *value function* for the last period:

$$V_T^i \left(\overline{x}_{T-1} \right) \equiv U^i \left(g_{iT} \left(\overline{x}_{T-1} \right), \sum_{j \neq i} g_{jT} \left(\overline{x}_{T-1} \right), M^i \left(x_i^*, \overline{x}_{T-1} \right) \right).$$

Moving backward to the before last period, each agent solves:

$$\max_{x_{iT-1}} \left\{ \begin{array}{c} w_i - x_{iT-1} + \beta_i \left(x_{iT-1} + x_{-iT-1} \right) - v_i \left(x_{iT-1} - M^i \left(x_i^*, \overline{x}_{T-2} \right) \right) \\ + \sigma V_T^i \left(\frac{x_{iT-1} + x_{-iT-1}}{n} \right) \end{array} \right\}$$

The optimal decision x_{iT-1} cancels out the addition of several marginal effects:

i) as when agents are myopic, in the current period:

$$-1 + \beta_i - v_i' \left(x_{iT-1} - M^i \left(x_i^*, \overline{x}_{T-2} \right) \right),$$

ii) but unlike the case of myopic agents, there is also a marginal effect on the next period $\sigma \frac{\partial}{x_{iT-1}} V_T^i \left(\overline{x}_{T-1} \right)$. This expression can be developed as follows:

$$\begin{aligned} \sigma \frac{\partial}{x_{iT-1}} V_T^i \left(\overline{x}_{T-1} \right) &= \sigma \left[\begin{array}{c} -g_{iT}' + \beta_i \left(g_{iT}' + \sum_{j \neq i} g_{jT}' \right) \\ -v_i' \left(x_{iT} - M^i \left(x_i^*, \overline{x}_{T-1} \right) \right) \left(g_{iT}' - M_2^i \left(x_i^*, \overline{x}_{T-1} \right) \right) \end{array} \right] \frac{\partial \overline{x}_{T-1}}{\partial x_{iT-1}}, \\ &= \sigma \left[-g_{iT}' + \beta_i \left(g_{iT}' + \sum_{j \neq i} g_{jT}' \right) \right] \frac{1}{n}, \\ &= \frac{\sigma}{n} \left[\left(-1 + \beta \right) \theta_i + \beta \sum_{j \neq i} \theta_j \right]. \end{aligned}$$

This second marginal effect explains the difference between the myopic and farsighted behaviors. Note that this difference owes nothing to the next period deviation from the moral motivation. Indeed, a marginal increase of x_{iT-1} has an impact on the next period moral motivation equal to $\theta_i dx_{iT-1}$ but this increase is exactly offset by the next period optimal contribution, as noticed from (4), leaving the gap $x_{iT} - M^i \left(x_i^*, \overline{x}_{T-1} \right)$ unchanged. The difference only goes through the effect on the last period self-centered part $w_i - g_{iT} \left(\overline{x}_{T-1} \right) + \beta_i \left(g_{iT} \left(\overline{x}_{T-1} \right) + \sum_{j \neq i} g_{-iT} \left(\overline{x}_{T-1} \right) \right)$.

Overall, the first order conditions are:

$$-1 + \beta - v_i' \left(x_{iT-1} - M^i \left(x_i^*, \overline{x}_{T-2} \right) \right) + \frac{\sigma}{n} \left[\left(-1 + \beta \right) \theta_i + \beta \sum_{j \neq i} \theta_j \right] = 0, \quad \forall i.$$

Hence, the best (dominant) response for each agent is:

$$\begin{aligned} x_{iT-1} &= -\frac{(1-\beta)}{v_i} + \frac{\sigma}{v_i n}\left[(-1+\beta)\,\theta_i + \beta\sum_{j\neq i}\theta_j\right] + M^i\left(x_i^*, \overline{x}_{T-2}\right) \ , \\ &\equiv g_{iT-1}\left(\overline{x}_{T-2}\right) \ . \end{aligned} \tag{5}$$

Observe again that:

$$g'_{iT-1} = M_2^i\left(.,.\right) = \theta_i.$$

Each agent's value function for the before last period is then:

$$\begin{aligned} V_{T-1}^i\left(\overline{x}_{T-2}\right) \equiv\ & U^i\left(g_{iT-1}\left(\overline{x}_{T-2}\right), \sum_{j\neq i} g_{jT-1}\left(\overline{x}_{T-2}\right), M^i\left(x_i^*, \overline{x}_{T-2}\right)\right) \\ & +\sigma V_T^i\left(\frac{\sum_{h=1}^{n} g_{hT-1}\left(\overline{x}_{T-2}\right)}{n}\right) \end{aligned}$$

Recursively it is possible to construct the agents' value functions for each date. There is no conceptual difficulty in this exercise but it is tedious and relegated to the Appendix. The important piece of information is that the individual problems at date t generically read as:

$$\max_{x_{iT-t}} \left\{ \begin{array}{c} w_i - x_{iT-t} + \beta_i\left(x_{iT-t} + x_{-iT-t}\right) - v_i\left(x_{iT-t} - M^i\left(x_i^*, \overline{x}_{T-t-1}\right)\right) \\ +\sigma V_{T-t+1}^i\left(\frac{x_{iT-t}+x_{-iT-t}}{n}\right) \end{array} \right\} .$$

And, using the notation $\Lambda = \frac{\sigma}{n}\sum_{h=1}^{n}\theta_h = \sigma\overline{\theta} < 1$, the dominant strategy t periods before the last can be written generically:

$$\begin{aligned} x_{iT-t} &= -\frac{(1-\beta)}{v_i} + \frac{\sigma}{v_i n}\left[(-1+\beta)\,\theta_i + \beta\sum_{j\neq i}\theta_j\right]\sum_{h=0}^{t-1}\Lambda^h + M^i\left(x_i^*, \overline{x}_{T-t-1}\right), \\ &\equiv g_{iT-t}\left(\overline{x}_{T-t-1}\right) \ . \end{aligned} \tag{6}$$

In particular, at the first period equilibrium decisions are:

$$\begin{aligned} x_{i0} &= -\frac{(1-\beta)}{v_i} + \frac{\sigma}{v_i n}\left[(-1+\beta)\,\theta_i + \beta\sum_{j\neq i}\theta_j\right]\sum_{h=0}^{T-1}\Lambda^h + x_i^* \ , \\ &= g_{i0}\left(x_i^*\right) \ . \end{aligned} \tag{7}$$

We are now in a position to investigate whether those contributions could decline over time.

Assumption 8 *Assume that:*

$$\begin{aligned} -\frac{(1-\beta)}{v_i}+\frac{\sigma}{v_i n}\left[(-1+\beta)\,\theta_i+\beta\sum_{j\neq i}\theta_j\right]\sum_{h=0}^{T-1}\Lambda^h &\leq 0, \\ (-1+\beta)\,\theta_i+\beta\sum_{j\neq i}\theta_j &\geq 0. \end{aligned}$$

Those two conditions on parameters are met for instance when agents value sufficiently the public good (β is large enough) and discount heavily the future (σ small enough). It can then be established:

Theorem 2 *Under assumption 8, the MPE is characterized by non increasing contributions over time.*

Proof. Observe first that $x_{i0}\leq x_i^*\ \forall i$, by the first inequality in Assumption A_8, hence $M^i\left(x_i^*,\overline{x}_0\right)\leq x_i^*\ \forall i$. Then, we also have:

$$x_{i0}-x_{i1}=\frac{\sigma}{v_i n}\left[(-1+\beta)\,\theta_i+\beta\sum_{j\neq i}\theta_j\right]\Lambda^{T-1}+\left(x_i^*-M^i\left(x_i^*,\overline{x}_0\right)\right)\ \geq 0,\ \forall i,$$

since, by the second inequality in 8 the first term in the right hand side of the above expression is positive and, as seen above $x_i^*-M^i\left(x_i^*,\overline{x}_0\right)\geq 0$. Repeating the comparison of successive contributions, one immediately sees that x_{it} is non increasing over time. ■

From (7), when the horizon growths large ($T\to\infty$), this first equilibrium decisions tend to:

$$x_{i0}=-\frac{(1-\beta)}{v_i}+\frac{\sigma}{v_i n}\left[(-1+\beta)\,\theta_i+\beta\sum_{j\neq i}\theta_j\right]\frac{1}{1-\Lambda}+x_i^*.$$

Similarly, considering that period t is the first one, and letting the time horizon go to infinity, it is easy to see that the dominant strategies become stationary feedback rules:

$$\begin{aligned} x_{it} &= -\frac{(1-\beta)}{v_i}+\frac{\sigma}{v_i n}\left[(-1+\beta)\,\theta_i+\beta\sum_{j\neq i}\theta_j\right]\frac{1}{1-\Lambda}+M^i\left(x_i^*,\overline{x}_{t-1}\right), \qquad (8) \\ &\equiv g_i\left(\overline{x}_{t-1}\right)\ ,\ \forall i. \end{aligned}$$

Clearly, under Assumption 8, the property of non increasing contributions carries over to the case of an infinite horizon.

A last question is in order: could this logic of behaviors explains the so-called restart effect? The restart effect can be understood in two related yet distinct ways. In the formalism of our model, when the condition of declining contributions is met, one would speak of a restart effect in either of the two following situations:

1. say that the duration of the game is first announced to be of $T/2$ periods, then at date $T/2-1$ there is a surprise restart announcement, according to which agents will play a further $T/2$ periods after date $T/2$. there is a restart effect if the contributions at date $T/2$ with the restart announcement, $x'_{i\frac{T}{2}}$, are larger than the contributions at the same date without the announcement, $x_{i\frac{T}{2}}$.

2. imagine now that the announcement is made at date $T/2$; there is a restart effect if the the contributions at date $T/2+1$ with the restart announcement, $x'_{i\frac{T}{2}+1}$, are larger than the contributions at date $T/2$ without the announcement, $x_{i\frac{T}{2}}$.

Let us investigate each possibility in turn. Without the surprise restart announcement, according to (6) or (5) at date $T/2$ agents would contribute:

$$x_{i\frac{T}{2}} = -\frac{(1-\beta)}{v_i} + \frac{\sigma}{v_i n}\left[(-1+\beta)\,\theta_i + \beta\sum_{j\neq i}\theta_j\right] + M^i\left(x_i^*, \overline{x}_{\frac{T}{2}-1}\right) .$$

With the restart announcement made at date $T/2-1$, agents treat the problem as if they were engaged in a new $\left(\frac{T}{2}+1\right)-$period game, with an initial moral motivation $M^i\left(x_i^*, \overline{x}_{\frac{T}{2}-1}\right)$, so their contribution in the next period of this new sequence (at date $T/2$) is going to be:

$$x'_{i\frac{T}{2}} = -\frac{(1-\beta)}{v_i} + \frac{\sigma}{v_i n}\left[(-1+\beta)\,\theta_i + \beta\sum_{j\neq i}\theta_j\right]\sum_{h=0}^{\frac{T}{2}}\Lambda^h + M^i\left(x_i^*, \overline{x}_{\frac{T}{2}-1}\right).$$

Comparing the contributions of the two scenarios at period $T/2$, one finds:

$$x'_{i\frac{T}{2}} - x_{i\frac{T}{2}} = \frac{\sigma}{v_i n}\left[(-1+\beta)\,\theta_i + \beta\sum_{j\neq i}\theta_j\right]\sum_{h=1}^{\frac{T}{2}}\Lambda^h \geq 0,$$

where the positive sign of the right-hand side is guaranteed from the second inequality in Assumption A_8. So,

Proposition 4 *A restart effect of the first kind occurs under Assumption 8.*

As for the second kind of restart effect, notice that the contributions in period $T/2+1$, after the announcement at date $T/2$, are going to be:

$$x'_{i\frac{T}{2}+1} = -\frac{(1-\beta)}{v_i} + \frac{\sigma}{v_i n}\left[(-1+\beta)\,\theta_i + \beta\sum_{j\neq i}\theta_j\right]\sum_{h=0}^{\frac{T}{2}-1}\Lambda^h + M^i\left(x_i^*, \overline{x}_{\frac{T}{2}}\right).$$

Next consider the difference:

$$\begin{aligned} x'_{i\frac{T}{2}+1} - x_{i\frac{T}{2}} &= \frac{\sigma}{v_i n}\left[(-1+\beta)\,\theta_i + \beta\sum_{j\neq i}\theta_j\right]\sum_{h=1}^{\frac{T}{2}-1}\Lambda^h \\ &\quad + M^i\left(x_i^*, \overline{x}_{\frac{T}{2}}\right) - \left[+M^i\left(x_i^*, \overline{x}_{\frac{T}{2}-1}\right)\right], \\ &= \frac{\sigma}{v_i n}\left[(-1+\beta)\,\theta_i + \beta\sum_{j\neq i}\theta_j\right]\sum_{h=1}^{\frac{T}{2}-1}\Lambda^h \\ &\quad + \theta_i\left[\overline{x}_{\frac{T}{2}} - \overline{x}_{\frac{T}{2}-1}\right]. \end{aligned}$$

Hence

$$x'_{i\frac{T}{2}+1} > x_{i\frac{T}{2}} \;\Leftrightarrow \frac{\sigma}{\theta_i v_i n}\left[(-1+\beta)\,\theta_i + \beta\sum_{j\neq i}\theta_j\right]\sum_{h=1}^{\frac{T}{2}-1}\Lambda^h \;> \overline{x}_{\frac{T}{2}-1} - \overline{x}_{\frac{T}{2}}\,.$$

Observe that, under Assumption 8, the right hand side of the above inequality is positive (because contributions are decreasing) and bounded. Indeed, the average contribution necessarily falls in the interval $[0, w_i]$, therefore $0 \leq \overline{x}_{\frac{T}{2}-1} - \overline{x}_{\frac{T}{2}} \leq w_i$. By contrast, if v_i approaches zero, the left hand side of the inequality tends to infinity. Thus:

Proposition 5 *Under Assumption 8, there exists values of parameters v_i of the weak moral motivation functions such that a restart effect of the second kind occurs.*

5 Discussion

Three attempts to explain the decline in contributions to public goods are closer to ours than any others. These are Kandori (2002), Klumpp (2005) and Ambrus & Pathak (2007).

The model proposed by Kandori (2002) is based on the general idea that the decline is generated by the erosion of norm and morale. This idea is very close to our concept of weak moral motivation , although the author takes a radically different modeling approach. As in other models, including our own, Kandori assumes that an individual's utility has two components: a material and a psychological one. The psychological component depends on the difference between his own effort (or contribution) and the median effort in the population. The dynamics is introduced in two ways. First, individuals revise their current effort with respect to the previous period median, which determines therefore a new current median. Second, in each period there is a small probability (the mutation rate) for each player to change his behavior. He ends up choosing randomly, i.e. each level of effort is chosen with equal probability. The author shows that the evolutionary stable equilibria are characterized by declining median effort. With respect to the experimental findings about the decay, this model does not account for variability in individual contributions since only the symmetric players' case is discussed. Furthermore, the model does not account for belief-based reciprocity , since players are assumed to behave myopically. Finally, the hypothesis that players adjust their effort with respect to the median effort, requires that each player observes all other players' individual efforts, a context that does not fit most of the available experimental data on voluntary contributions.

Ambrus & Pathak (2007) consider a mixed population of players, which consists of selfish and reciprocal types. While the selfish players have homogenous preferences, there is heterogeneity among reciprocal types, which is captured by reciprocity functions. The dynamics of the model is generated by the behavior of selfish players who have an incentive to contribute large amounts in early periods because of their influence on future contributions of reciprocal types. The incentive to cooperate of selfish players depends on the number of remaining periods in the repeated game. As the end of the game approaches,

the selfish players switch to their Nash contribution of the one-shot game, i.e. zero contribution. The authors assume a continuous strategy space (players can contribute any real number between 0 and 1). The key assumption for the decline is common knowledge of preferences of all players, in particular reciprocity functions are common knowledge. Although such an assumption can have some realism in a population of players who know each other well and have experienced frequent interactions over a long period, it does not apply to most experimental data, where subjects are anonymous and interact only for a few periods. Furthermore, the decay in average contribution is obtained by a decline in individual contributions for both types. This requirement seems unnecessarily strong, and does not match individual behavior in voluntary contribution experiments. The experimental data reveals a high variability of individual contributions from period to period (see e.g. Keser & Van Winden, 2000), which is typically not captured in their model. In contrast, our model does not require common knowledge and allows both for increasing and decreasing individual contributions, with the weaker requirement that aggregate contributions cannot be larger than the initial aggregate moral motivation.

In Klumpp (2005) players are endowed with social preferences. Their utility representation has two additively separable components : material utility and psychological utility. The stage game admits two symmetric Nash equilibria : one where no player contributes to the public good, and one where each player contributes a strictly positive amount. While this dynamic game admits multiple Nash equilibria, the author shows that there is a unique maximal symmetric equilibrium path in pure strategies , for which individual contributions decline. As in Ambrus & Pathak (2007) restricting attention to this particular path is too strong for generating a decline in average contributions, and does not correspond to most available data. More important however, is that the temporal profile of the maximal equilibrium path is not compatible with the pattern typically observed in most experiments on voluntary contributions to a linear public good. In Klumpp's model, the maximal equilibrium path is one where all players contribute all of their endowement up to some date, after which they start lowering their contribution down to a level that is approximately equal to zero. In contrast, average contributions in linear public goods

start at a level that is in between half the endowment and the Nash contribution (see Laury & Holt, 2008) and then declines slowly to reach a positive level that is significantly larger than the Nash contribution of the constituent game. In Klumpp's model the average contribution falls very sharply from 100% contribution to nearly 0% contribution over a few periods.

The conclusion of this discussion is that the models proposed by Kandori (2002), Klumpp (2005) and Ambrus & Pathak (2007) have the clear merit to indicate research directions that are worthwhile exploring. But like many pioneering contributions, they rely on unnecessarily strong assumptions to generate the decay of the average contribution and/or their outcomes fit only very roughly with observed data. By contrast, our explanation does not rest on the assumption of common knowledge; it does not depend either on an equilibrium selection argument or on a evolutionary game concept. And the generated outcome accounts for most, if not all, of the empirical regularities observed in the experimental literature.

6 Conclusion

Our aim in this paper was to provide a general framework that accounts for the decay of the average contribution observed in most experiments on voluntary contributions to a public good. Each player balances her material utility loss from contributing with her psychological utility loss of deviating from her moral ideal. The central idea of our model is that people's moral motivation is "weak": their judgement about what is the right contribution to a public good can evolve in the course of interactions, depending partly on observed contributions and partly on an intrinsic "moral ideal". The decline of the average contribution is generated by two effects, that presumably can be in conflict. The first one is a downward or selfish-bias due to the presence of material payoffs in agents' preferences. The unambiguous outcome is that each player contributes a little less than her moral ideal. The second - and novel - effect is that, because individuals' moral ideals

are weak, in the sense defined above, the upward bias induced by moral motivations can become less stringent from one period to the other.

We started by showing that if players behave myopically, *i.e.* they do not take into account the influence of their contribution on others' future moral motivations, under natural conditions the average contribution is non-increasing as the contribution game is repeated.

The hypothesis of myopic behavior seems to us the most appealing one with respect to the experimental data about subjects' behavior. However, we cannot preclude the fact that some subjects act as farsighted players and try to manipulate others' moral motivation. Therefore, we provided an extension of our basic model, to account for the more general case of farsighted players. Assuming farsighted behavior, we showed that the decline arises if agents value sufficiently the public good (β is large enough) and discount sufficiently the future (σ small enough). The requirement, that under farsighted behavior the discount rate must be small with respect to the value of the public good, justifies our preference for the simpler assumption of myopic behavior for accounting for the decline.

The proposed framework allows for heterogeneity in players' endowment, preferences, and moral ideal. It therefore encompasses a huge variety of individual behaviors. It predicts many observed experimental regularities[12]: over-contributions, heterogeneity of contributions, declining average contributions, final over-contributions and the restart effect.

The proposed model can be interpreted from two, apparently different, behavioral angles: reciprocity and moral motivation. While the reciprocity motive has been widely documented in the recent theoretical and empirical economic literature, the moral motivation hypothesis has not retained much attention. There is a large historical trend in the philosophical literature concerned with moral motivation and the strength of will, that is relevant with respect to our assumptions. In our model there is a strong link between these two dimensions. The reason is that reciprocal behavior is somehow grounded on an internal deliberation process, through which individuals combine their intrinsic motivation to contribute with external pressures in their environment. The hypothesis of weak moral

[12] Actually we are aware of no such regularities that the model does not predict.

moral can therefore be thought as a means to rationalize reciprocal behavior. We believe that such hypothesis can be tested with a carefully designed experiment.

As a final thought, our model is also related to the endogenous preferences literature (see Bowles, 1998). We believe that, in contrast to models based on heterogeneity of player- types, models based on endogenous preferences are well-adapted to account for various aspects of the behavioral patterns observed in experiments where subjects interact repeatedly.

Appendix

A Derivation of the Markov Perfect Equilibrium

In the text, equilibrium decisions for periods T and $T-1$ have been given. Moving backward to period $T-2$, each agent's decision solves:

$$\max_{x_{iT-2}} \left\{ \begin{array}{c} w_i - x_{iT-2} + \beta_i \left(x_{iT-2} + x_{-iT-2} \right) - v_i \left(x_{iT-2} - M^i \left(x_i^*, \overline{x}_{T-3} \right) \right) \\ +\sigma V_{T-1}^i \left(\frac{x_{iT-2} + x_{-iT-2}}{n} \right) \end{array} \right\} .$$

The marginal effects of changing x_{iT-2} are now as follows:

i) As before there are effects on the current utility:

$$-1 + \beta_i - v_i' \left(x_{iT-2} - M^i \left(x_i^*, \overline{x}_{T-3} \right) \right),$$

ii) there are also marginal effects on the discounted indirect utility of period $T-1$:

$$\begin{aligned} & \sigma \left[\begin{array}{c} -g_{iT-1}' + \beta_i \left(g_{iT-1}' + \sum_{j \neq i} g_{jT-1}' \right) \\ -v_i' \left(x_{iT-1} - M^i \left(x_i^*, \overline{x}_{T-2} \right) \right) \left(g_{iT-1}' - M_2^i \left(x_i^*, \overline{x}_{T-2} \right) \right) \end{array} \right] \frac{\partial \overline{x}_{T-2}}{\partial x_{iT-2}} , \\ = & \frac{\sigma}{n} \left[\left(-1 + \beta \right) \theta_i + \beta \sum_{j \neq i} \theta_j \right] , \end{aligned}$$

iii) and finally, there are marginal effects on discounted indirect utility of period T:

$$\begin{aligned}
& \sigma^2 \left[\begin{array}{c} -g'_{iT} + \beta_i \left(g'_{iT} + \sum_{j \neq i} g'_{jT} \right) \\ -v'_i \left(x_{iT} - M^i \left(x_i^*, \overline{x}_{T-2} \right) \right) \left(g'_{iT} - M_2^i \left(x_i^*, \overline{x}_{T-1} \right) \right) \end{array} \right] \frac{\partial \overline{x}_{T-1}}{\partial x_{iT-2}} \\
= \; & \sigma^2 \left[\left(-1 + \beta \right) \theta_i + \beta \sum_{j \neq i} \theta_j \right] \frac{\sum_{h=1}^n g'_{hT-1}}{n} * \frac{\partial \overline{x}_{T-2}}{\partial x_{iT-2}} \; , \\
= \; & \frac{\sigma^2}{n^2} \left[\left(-1 + \beta \right) \theta_i + \beta \sum_{j \neq i} \theta_j \right] \sum_{h=1}^n \theta_h
\end{aligned}$$

The first order condition is therefore:

$$-1 + \beta_i - v'_i \left(x_{iT-2} - M^i \left(x_i^*, \overline{x}_{T-3} \right) \right) + \frac{\sigma}{n} \left[\left(-1 + \beta \right) \theta_i + \beta \sum_{j \neq i} \theta_j \right] \left(1 + \frac{\sigma}{n} \sum_{h=1}^n \theta_h \right) = 0 \, .$$

And the dominant response can be expressed again as:

$$\begin{aligned}
x_{iT-2} &= -\frac{\left(1 - \beta \right)}{v_i} + \frac{\sigma}{vn} \left[\left(-1 + \beta \right) \frac{\theta_i}{n} + \beta \sum_{j \neq i} \frac{\theta_j}{n} \right] \left(1 + \frac{\sigma}{n} \sum_{h=1}^n \theta_h \right) + M^i \left(x_i^*, \overline{x}_{T-3} \right) \; , \\
&\equiv g_{iT-2} \left(\overline{x}_{T-3} \right) \; .
\end{aligned}$$

The marginal effects of changing x_{iT-3} are now as follows:

i) the effects on the current utility are:

$$-1 + \beta_i - v'_i \left(x_{iT-3} - M^i \left(x_i^*, \overline{x}_{T-4} \right) \right) ,$$

ii) there are also marginal effects on the discounted indirect utility of period $T-2$:

$$\begin{aligned}
& \sigma \left[\begin{array}{c} -g'_{iT-2} + \beta_i \left(g'_{iT-2} + \sum_{j \neq i} g'_{jT-2} \right) \\ -v'_i \left(x_{iT-2} - M^i \left(x_i^*, \overline{x}_{T-3} \right) \right) \left(g'_{iT-2} - M_2^i \left(x_i^*, \overline{x}_{T-3} \right) \right) \end{array} \right] \frac{\partial \overline{x}_{T-3}}{\partial x_{iT-3}} \; , \\
= \; & \frac{\sigma}{n} \left[\left(-1 + \beta \right) \theta_i + \beta \sum_{j \neq i} \theta_j \right] \; ,
\end{aligned}$$

iii) and also, there are marginal effects on discounted indirect utility of period $T-1$:

$$\begin{aligned}
& \sigma^2 \begin{bmatrix} -g'_{iT-1} + \beta_i \left(g'_{iT-1} + \sum_{j \neq i} g'_{jT-1} \right) \\ -v'_i \left(x_{iT-1} - M^i \left(x_i^*, \overline{x}_{T-2} \right) \right) \left(g'_{iT-1} - M_2^i \left(x_i^*, \overline{x}_{T-2} \right) \right) \end{bmatrix} \frac{\partial \overline{x}_{T-2}}{\partial x_{iT-3}} \\
= \; & \sigma^2 \left[(-1+\beta)\,\theta_i + \beta \sum_{j \neq i} \theta_j \right] \frac{\sum_{h=1}^n g'_{hT-2}}{n} * \frac{\partial \overline{x}_{T-3}}{\partial x_{iT-3}} \; , \\
= \; & \frac{\sigma^2}{n^2} \left[(-1+\beta)\,\theta_i + \beta \sum_{j \neq i} \theta_j \right] \sum_{h=1}^n \theta_h
\end{aligned}$$

iv) and finally, there are marginal effects on discounted indirect utility of period T:

$$\begin{aligned}
& \sigma^3 \begin{bmatrix} -g'_{iT} + \beta_i \left(g'_{iT} + \sum_{j \neq i} g'_{jT} \right) \\ -v'_i \left(x_{iT} - M^i \left(x_i^*, \overline{x}_{T-1} \right) \right) \left(g'_{iT} - M_2^i \left(x_i^*, \overline{x}_T \right) \right) \end{bmatrix} \frac{\partial \overline{x}_{T-1}}{\partial x_{iT-3}} \\
= \; & \sigma^2 \left[(-1+\beta)\,\theta_i + \beta \sum_{j \neq i} \theta_j \right] \frac{\sum_{h=1}^n g'_{hT-3}}{n} * \frac{\sum_{h=1}^n g'_{hT-2}}{n} * \frac{\partial \overline{x}_{T-3}}{\partial x_{iT-3}} \; , \\
= \; & \frac{\sigma^2}{n^3} \left[(-1+\beta)\,\theta_i + \beta \sum_{j \neq i} \theta_j \right] \left(\sum_{h=1}^n \theta_h \right)^2
\end{aligned}$$

The first order condition is therefore:

$$-1+\beta_i - v'_i \left(x_{iT-3} - M^i \left(x_i^*, \overline{x}_{T-4} \right) \right) + \frac{\sigma}{n} \left[(-1+\beta)\,\theta_i + \beta \sum_{j \neq i} \theta_j \right] \left[1 + \frac{\sigma}{n} \sum_{h=1}^n \theta_h + \frac{\sigma^2}{n^2} \left(\sum_{h=1}^n \theta_h \right)^2 \right] =$$

And the dominant response can be expressed again as:

$$\begin{aligned}
x_{iT-3} &= -\frac{(1-\beta)}{v_i} + \frac{\sigma}{v_i n} \left[(-1+\beta)\,\theta_i + \beta \sum_{j \neq i} \theta_j \right] \left[1 + \frac{\sigma}{n} \sum_{h=1}^n \theta_h + \frac{\sigma^2}{n^2} \left(\sum_{h=1}^n \theta_h \right)^2 \right] + M^i \left(x_i^*, \overline{x}_{T-4} \right) \; , \\
&\equiv g_{iT-2} \left(\overline{x}_{T-3} \right) \; .
\end{aligned}$$

Repeating the logic, and using the notation $\Lambda = \frac{\sigma}{n} \sum_{h=1}^n \theta_h$, the dominant strategy t periods before the last can be written generically:

$$\begin{aligned}
x_{iT-t} &= -\frac{(1-\beta)}{v_i} + \frac{\sigma}{v_i n} \left[(-1+\beta)\,\theta_i + \beta \sum_{j \neq i} \theta_j \right] \left[1 + \Lambda + \Lambda^2 + ... + \Lambda^{t-1} \right] + M^i \left(x_i^*, \overline{x}_{T-t-1} \right) \; , \\
&= -\frac{(1-\beta)}{v_i} + \frac{\sigma}{v_i n} \left[(-1+\beta)\,\theta_i + \beta \sum_{j \neq i} \theta_j \right] \sum_{h=0}^{t-1} \Lambda^h + M^i \left(x_i^*, \overline{x}_{T-t-1} \right) \; , \\
&\equiv g_{iT-t} \left(\overline{x}_{T-t-1} \right) \; .
\end{aligned}$$

References

[1] Anderson S., Goeree J., Holt C. (1998), "A theoretical analysis of altruism and decision error in public goods games", Journal of Public Economics, 70, 297–323.

[2] Anderson S., Goeree J., Holt C. (2004), "Noisy directional learning and the logit equilibrium", Scandinavian Journal of Economics, (106) September 2004, 581-602.

[3] Andreoni J., (1988), "Why free ride ? Strategies and learning in public goods experiments", Journal of Public Economics, 37, 291-304.

[4] Andreoni, J., (1988), "Privately provided public goods in a large economy : The limits of altruism", Journal of Public Economics 35, 57–73.

[5] Andreoni J., (1990), "Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving", Economic Journal, 100, 464-477.

[6] Andreoni, J., (1995), "Cooperation in Public-Goods Experiments - Kindness or Confusion?" American Economic Review, 85(4): 891-904.

[7] Andreoni, J., Miller, J. H. (1996), Giving According to GARP: An Experimental Study of Rationality and Altruism, Working Paper, University of Wisconsin, Madison.

[8] Andreoni J., Croson R. (2002), "Partners versus Strangers: Random Rematching in Public Goods Experiments", in Handbook of Experimental Economics Results, C. R. Plott and V. L. Smith (eds)

[9] Bolton, G. E., Ockenfels, A., (2000). "ERC: A Theory of Equity, Reciprocity, and Competition." American Economic Review, 90(1), 166-93.

[10] Bowles, S., (1998), "Endogenous Preferences: The Cultural Consequences of Markets and Other Economic Institutions", Journal of Economic Literature, Vol. XXXVI, 75-111.

[11] Brandts, J., Schram, A. (2001), Cooperative Gains or Noise in Public Goods Experiments: Applying the Contribution Function Approach, Journal of Public Economics 79, 399–427.

[12] Brekke K., Kverndokk S., Nyborg K. (2003), "An economic model of moral motivation", Journal of Public Economics 87:1967-1983.

[13] Carpenter, J. (2007). "Punishing Free-Riders: How Group Size Affects Mutual Monitoring and the Provision of Public Goods." Games and Economic Behavior, 60(1): 31-51.

[14] Croson, R., (1996). Partners and strangers revisited. Economics Letters 53, 25–32.

[15] Croson, R. (2007), "Theories of commitment, altruism & reciprocity: evidence from linear public goods games". Economic Inquiry, 45, 199-216.

[16] Croson, R. T. A. (2000). Thinking like a game theorist: factors affecting the frequency of equilibrium play. Journal of Economic Behavior and Organization 41: 299-314.

[17] Croson, R. T. A., Fatas, E., & Neugebauer, T. (2005). Reciprocity, matching & conditional cooperation in two public goods games. Economics Letters, 87, 95-102.

[18] Davis, D. et Holt C. (1993), Experimental economics, Princeton University Press, Princeton, NJ.

[19] Dawes, R. M. and Thaler, R. H. (1988), Anomalies: Cooperation, Journal of Economic Perspectives 2, 187-197.

[20] Drèze, J.H., and D. De la Vallee Poussin (1971), "A tâtonnement process for public goods", *Review of Economic Studies*, 38, 133-50.

[21] Fehr, E., Schmidt K. (1999), "A theory of fairness, competition and cooperation", Quarterly Journal of Economics, 114, 817-868.

[22] Fehr, E., and Gächter, S. (2000). "Cooperation and Punishment in Public Goods Experiments." American Economic Review, 90(4): 980-94.

[23] Fischbacher U., Gächter S. (2009), "Social Preferences, Beliefs and the Dynamics of Free Riding in Public Good Experiments". American Economic Review, (forthcoming)

[24] Fischbacher, U., Gächter, S., & Fehr, E. (2001). Are people conditionally cooperative? Evidence from a public goods experiment. Economics Letters, 71, 397-404.

[25] Gächter, S., and Fehr, E. (1999). "Collective Action as a Social Exchange." Journal of Economic Behavior & Organization, 39(4): 341-69.

[26] Harsanyi J. (1980), "Rule Utilitarianism, Rights, Obligations and the Theory of Rational Behavior", Theory and Decision, 12, 115-133.

[27] Herrmann, B., Thöni, C., and Gächter, S. (2008). "Antisocial Punishment across Societies." Science, 319: 1362-67.

[28] Holton, R., (2003), "How is Strength of Will Possible?" in S. Stroud and C Tappolet (eds.) Weakness of Will and Practical Irrationality (Oxford: Clarendon Press, 2003), pp. 39-67

[29] Holt C., Laury S. (2008), "Theoretical explanations of treatment effects in voluntary contributions experiments", in Hanbook of experimental Economics Results (vol. 1), C. Plott & V. Smith (eds), Noth-Holland, 846-855.

[30] Houser D., Kurzban (2002), "Revisiting kindness and confusion in public good experiments", American Economic Review, 92, 4, 1062-1069.

[31] Isaac M., Walker J. (1988), "Nash as an Organizing Principle in the Voluntary Provision of Public Goods : Experimental Evidence", Experimental Economics, 1, 191-206.

[32] Isaac R.M., Walker J.M., Thomas S. H. (1984), « Divergent Evidence on Free Riding : An Experimental Examination of Possible Explanations », Public Choice, 43, 113-149.

[33] Isaac M., Walker J., Williams A. (1994), "Group size and the voluntary provision of public goods: Experimental evidence utilizing large groups", Journal of Public Economics 54, 1–36.

[34] Kandori M. (2002), "The erosion and sustainability of norms and morale", working paper, CTRJE-F-169, University of Tokyo.

[35] Keser, C. (1996), "Voluntary contributions to a public good when partial contribution is a dominant strategy", Economics Letters, 50, 359-366.

[36] Keser C., Van Winden F., 2000, "Conditional cooperation and voluntary contributions to public goods", Scandinavian Journal of Economics, 102, 1, 23-39.

[37] Kolm, S.-C. (1984), "Théorie de la réciprocité et du choix des systèmes économiques", Revue Economique, 35(5), 871-910.

[38] Kreps D., Milgrom P., Roberts J.,Wilson R., (1982), "Rational cooperation in the finitely repeated Prisoners' Dilemma." Journal of Economic Theory, 27, 1982,245-52.

[39] Laffont J.-J., (1975), "Macroeconomic Constraints, Economic Efficiency and Ethics : An Introduction to Kantian Economics", Economica, 42, 430-437.

[40] Laury S., Holt, C., (2008), "Voluntary Provision of Public Goods : Experimental Results with Interior Nash Equilibria", in Handbook of Experimental Economics Results, C. Plott & V. Smith (eds), Elsevier, 792-801.

[41] Ledyard, J. (1995), Public goods: A survey of experimental research, in J.H. Kagel and A.E. Roth (eds.), The handbook of experimental economics, Princeton University Press.

[42] Masclet, D., Noussair, C., Tucker, S., and Villeval, M. C. 2003. "Monetary and Non-monetary Punishment in the Voluntary Contributions Mechanism." American Economic Review, 93(1): 366-80.

[43] Neugebauer T., Perote J., Schmidt U., Loos M., (2009), "Selfish-biased conditional cooperation: On the decline of contributions in repeated public goods experiments". Journal of Economic Psychology, 30(1), 52-60.

[44] Nyborg, K. (2000), "Homo economicus and homo politicus: Interpretation and aggregation of environmental values.", Journal of Economic Behavior and Organization 42, 305–322.

[45] Palfrey, T., & Prisbey, J. (1997). Anomalous behavior in public goods experiments: how much and why. American Economic Review, 87 (5), 829-846.

[46] Rabin, M., (1993). "Incorporating Fairness into Game Theory and Economics." American Economic Review, 83, 1281-1302.

[47] Sefton M., Steinberg (1996), "Reward structures in public good experiments", Journal of Public Economics, 61, 263–287.

[48] Selten, R., Buchta, J. (1998), "Experimental sealed bid first price auctions with directly observed bid functions", in I. E. D. Budescu, I. Erev and R. Zwick, (eds.), Games and Human Behavior: Essays in Honor of Amnon Rapoport, Lawrence Erlbaum Associates,Mahwah, NJ.

[49] Sefton, M., Shupp, R., and Walker, J. M. (2007). "The Effect of Rewards and Sanctions in Provision of Public Goods Economic Inquiry, 45(4): 671–90.

[50] Sonnemans, J., Schram, A., Offerman, T., (1999). Strategic behavior in public goods games: when partners drift apart. Economics Letters 62, 35–41.

[51] Weimann, J. (1994), "Individual behavior in a free riding experiment", Journal of Public Economics 54:185-20.

[52] Zangwill, N., (2003), "Externalist moral motivation", American Philosophical Quarterly, 40(2), 143-154.

[53] Zhang X.,hu F., (2007), “Group Size and Incentive to Contribute: A Natural Experiment at Chinese Wikipedia”, working paper.