ELSEVIER

Journal of Economic Behavior & Organization
Vol. 55 (2004) 377–395

JOURNAL OF
Economic Behavior
& Organization

www.elsevier.com/locate/econbase

The price of a price: on the crowding out and in of social norms

Maarten C.W. Janssen [a,*], Ewa Mendys-Kamphorst [b]

[a] *Department of Economics, Erasmus University Rotterdam, Burg. Oudlaan 50, 3062 PA Rotterdam, The Netherlands*
[b] *SEOR-Erasmus Competition and Regulation Institute, Erasmus University Rotterdam, Burg. Oudlaan 50, 3062 PA Rotterdam, The Netherlands*

Received 2 July 2002; received in revised form 20 September 2002; accepted 27 November 2002
Available online 28 May 2004

Abstract

We study the impact of financial incentives on social approval, showing that in a society with altruists and egoists, who all care about social approval, introducing financial incentives to agents to contribute to a socially desirable outcome may actually decrease the number of contributions. Withdrawing the financial incentive does not restore the norm to contribute and may reduce the level of contributions even further. When the norm has disappeared, it may be possible to restore voluntary contributions by first introducing high price and then reducing it, but such an operation is costly and its success uncertain.

JEL classification: I18; D10; Z13

Keywords: Social norms; Intrinsic and extrinsic motivation; Network effects

1. Introduction

According to most economic theory, if people are willing to make an effort even if they are not financially compensated, they should be more eager to make this effort if they get paid. The underlying assumption is that existing non-financial motivation is unaffected when a financial reward is introduced. Hence, different kinds of financial and non-financial motivations can be added up. This assumption allows economists to treat the issue of non-financial incentives as being a matter of exogenous personal preferences, that cannot

* Correspondig author.
E-mail addresses: janssen@few.eur.nl (M.C.W. Janssen), mendys@few.eur.nl (E. Mendys-Kamphorst).

0167-2681/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.jebo.2002.11.004

be affected by economic policies. In this way economics is able to reduce the problem of motivating people to designing optimal financial compensation schemes.

This approach aroused occasional discomfort among some representatives of the economic profession who argued that *homo economicus* and *homo sociologicus* could not be so easily separated. People are usually motivated by a combination of forces that may reinforce or weaken one another. One of the first arguments against basing government policies on economic incentives alone was provided by Richard Titmuss (1970). After comparing the American (mostly paid or providing other benefits) and British (entirely voluntary and unpaid) systems of obtaining blood for medical purposes, he concluded that the paid system results in shortages and a lower quality of blood supply. He also noticed that the social characteristics of contributors in Britain differed from the characteristics of blood contributors in the US. In Britain they were representative of the population, while in America they tended to have lower income and lower education. His conclusion was that paying for blood destroys an altruistic motivation to contribute. Moreover, he claimed that this motivation is destroyed permanently, and that removing the monetary incentive would not restore the altruistic motivation, at least not soon.

Although Titmuss's evidence did not prove convincing to all (see e.g. Arrow, 1972), in recent years many experiments[1] have been conducted demonstrating the importance of rethinking the interaction between different types of motivation. Among other things, it has been observed that monetary incentives can "crowd out" other sorts of motivation, often called "intrinsic" motivation. Once crowded out, the "intrinsic" motivation often does not come back after the monetary incentive has been removed. This is true, for example, for a field study conducted in a group of day-care centers in Israel (Gneezy and Rustichini, 2000a) where parents arrived late to collect their children. After introducing a monetary fine for late-coming parents, more parents began coming late. Removing the fine did not restore the initial situation. Frey and Götte (1999) report that in another field experiment the amount of work done by volunteers fell after a financial incentive was introduced.

In this paper, we want to interpret the above-mentioned empirical results by analyzing one of the possible mechanisms underlying the interaction of financial and non-financial motivations. We analyze a situation in which people's actions are partly driven by the desire of social approval or status, which we label "social reward." Introducing financial incentives may eliminate or reduce the source of status, thereby reducing that type of motivation. The way we model social reward is related to a behavioral theory of crowding out presented by Dickinson (1989, p. 6). She writes: "In our society, people are often admired and praised when they engage in intrinsically controlled behavior (. . .). The very fact that the behavior is unrelated to any obvious extrinsic consequence is, in part, the basis for the approval". In such an environment, providing a financial incentive is likely to decrease social reward.

The importance of social recognition as a motivation source has been noted in a number of studies. Studying donations to charity, Harbaugh (1998a,b), Glazer and Konrad (1996), and Andreoni and Petrie (2000) note that charity organizations generally publish lists of contributors together with the (approximate) value of their contributions. The fact that people

[1] For an overview of psychological literature on this subject, see Deci (1999); for a survey of empirical evidence, see Frey and Jegen (2001); for economic experimental evidence, see e.g. Frey (1997), Frey and Götte (1999), Fehr and Gächter (2000) and Gneezy and Rustichini (2000a, b).

care about others seeing their contributions suggests that there is certain social recognition related to charitable actions. Andreoni and Petrie (2000), and Fehr and Gächter (1999) present experiments in which people contribute more to a public good if their contributions are known to others.

We analyze the interaction between financial and non-financial motivations and the resulting implications of introducing payment for performing a socially desirable task by means of a simple model. For brevity, we will speak of performing such a task as "contributing".

There are two types of individuals in the population: altruists and egoists.[2] Both types are motivated to the same extent by extrinsic social reward, and contributing also involves the same cost to both types of people. In addition, both types value money, but not to the same degree: partly motivated by an intrinsic desire to help others altruists derive less utility from money than egoists. Hence, egoists will optimally not contribute unless the sum of social and financial reward compensates them for the cost. Altruists, on the other hand, contribute if the sum of financial, social and intrinsic rewards exceeds the cost. We assume that the social reward increases with the number of altruistic and decreases with the number of egoistic contributors and that it is only in place when at least some altruists contribute. As a consequence, altruists impose positive, and egoists negative, externalities on other contributors.

To study the dynamics of social change, we use an evolutionary model in which each type can follow one of two available strategies: contribute or not contribute. For the most part, an individual who adjusts his strategy chooses the one that gives him a higher expected utility assuming that others will behave as in the previous period. However, at each moment there is also a small chance that people will decide against their own interest.[3]

Using this framework, we are able to make a distinction between medium-run and long-run equilibria. In the medium run, the dynamics of the system is driven by the best-reply deterministic dynamics. If two equilibria exist, either of them may emerge in the medium run, depending on the initial state. In the (ultra) long run, the fact that people sometimes behave irrationally matters. In general (even if there are multiple medium-run equilibria), there is a unique equilibrium that prevails in the long run and this is the equilibrium that is *stochastically stable*.

By means of this model we obtain the following main results:

(i) In the absence of financial incentives, the long-run equilibrium may involve all altruists contributing and no, or some, egoists. This long-run equilibrium will arise even if no one contributes initially.

[2] The presence of altruistic types can be justified both on evolutionary and empirical grounds. An explanation of why and how a concern for other people may survive evolutionary selection can be found in, e.g. Bester and Güth (1998). Frank (1987), Bar-Gill and Fershtman (2001), Stewart (1992), and Güth and Kliemt (2000) show that populations in which egoistic and altruistic types coexist may be evolutionary stable. In an empirical study, Harbaugh (1998b) concludes that charity donations are partly motivated by altruism.

[3] Note that this specification follows the (mainstream) evolutionary game theoretic literature (see Young, 1998, or Weibull, 1995). In this literature, preferences are given and the evolutionary (or learning) process takes place on strategies, not on preferences. In our model the preferences of both types are given. In the indirect evolutionary approach the evolutionary process selects those subjective utility functions that result in a higher objective payoff than others (see e.g. Bester and Güth, 1998, or Güth and Kliemt, 2000). As there is no obvious candidate for objective pay-off in our model, it is difficult to follow the indirect evolutionary approach.

(ii) Providing financial incentives to increase contributions may have adverse effects in the medium run when, as a result, the social norm to contribute is destroyed.
(iii) The effect of financial incentives on the level of contributions may be non-monotonic: contributions increase, then decrease, and finally increase again as a small, intermediate, respectively, high financial incentive is introduced.
(iv) Even if introducing a financial reward leads to an increase in contributions in the short and medium run, it may have adverse effects in the long run.
(v) Once a norm has been crowded out, it can sometimes be restored with a help of a financial reward (*crowded in*). However, reaching the original level of the social reward is uncertain.
(vi) Without crowding in, the norm takes a long time to reappear.

Below, we will briefly explain the main intuition for these findings. We start with no financial reward being offered for contributing. We are especially interested in situations with two types of medium-run equilibria: Either all altruists contribute, or none. In the first case, a large number of contributors ensures that the social norm to contribute exists, and together with the intrinsic motivation, the social reward is high enough to outweigh the cost of contributing. In the second case, no one contributes, and as a consequence there is no social norm to contribute. Under certain conditions, the equilibrium with a social norm to contribute is stochastically stable. Hence, if no financial incentive is offered for a long enough time, all altruists will eventually contribute, even if initially no one did it.

Given this starting position with all altruists contributing, we study the effect of introducing a financial reward for contributing. On one hand, a price increases the altruists' utility from contributing. On the other hand, if the price for contributing is high enough, this will induce (more) egoists to contribute as well, which decreases the incentives of altruists. If the negative externalities created by egoists are strong enough, the sum of intrinsic motivation, price and social reward may become too low to encourage altruists to contribute. This is the crowding out effect. More generally, the impact of a price on the contributions is non-monotonic. Our results resemble, in this respect the experimental findings of Gneezy and Rustichini (2000b). If the price is relatively small, the social reward decreases only a little, and altruists keep contributing. If the price is very high, it compensates the altruists for the decrease in the social reward, and they therefore continue contributing. If the price takes intermediate values, it will be too low to compensate for the decrease in social reward, so altruists will stop contributing and the social reward will fall to zero. With no altruists contributing, it may well be that egoists will also stop contributing in the medium run as the social reward is non-existent. As a result, contributions will fall to zero.

Even if altruists will not stop contributing in the medium run and the medium run supply increases, contributions may actually decrease in the long run. This happens when the equilibrium with contributors loses its stochastic stability property after a financial reward has been introduced. The intuition behind this result is that an increased number of egoistic contributors decreases the social reward and that the subsequent decrease in altruists' utility from contributing may not be compensated by the financial incentive. If the altruists' utility from contributing becomes small, it suffices for only a few altruists to stop contributing by mistake for the utility from donating to fall below zero, so that all the other altruists stop contributing as well.

Next, we analyze what happens if the reward is withdrawn, possibly because the authorities have realized its adverse consequences. If the social reward has not been crowded out but substantially weakened, so that the financial reward was needed to encourage altruists to contribute, removing the price will result in the equilibrium with no contributions. On the other hand, if the norm has not been weakened too much by the financial incentive, the altruists will keep contributing even after the price is removed.

If no altruists contribute, but the equilibrium with contributors is stochastically stable in the absence of a financial reward, the social norm will eventually be rebuilt, but it will take a very long time. Alternatively, the authorities can try to rebuild the norm by introducing a very high price that will attract both altruistic and egoistic contributors. This price can subsequently be reduced, since the presence of altruists is a source of utility for other contributors. Thus, it is always possible to "crowd in" some social reward. However, bringing it to a level that makes it sufficient to ensure contributions even without a financial incentive is more difficult. It turns out that precisely in those cases where a financial reward can crowd out the norm immediately and thus is most harmful, the effectiveness of crowding in a sufficiently large social reward is most limited. It may be successful, but it requires choosing the price sequence very carefully, and the final outcome remains uncertain. Thus, destroying a norm is easier and faster than rebuilding it. Moreover, since crowding in requires setting a very high price initially, it may be too costly to make it practically feasible.

The rest of the paper is organized as follows. Section 2 introduces the model. Section 3 analyzes the type of equilibria that may arise under voluntary contribution systems. Conditions are stated under which a social norm to contribute will emerge as a stochastically stable equilibrium in the long run. Section 4 studies the effects of introducing a financial reward. In Section 5, we examine the consequences of removing the financial reward. Section 6 discusses crowding in, and Section 7 concludes with a discussion of the role of the assumptions underlying the model that we use.

2. The model

Our society consists of two types of people, whom we simply call (for easy reference) altruists and egoists. The total number of each type is denoted by $\bar{N}_a$ and $\bar{N}_e$, respectively, and N_a^t and N_e^t denote the number of individuals of each type who contribute in period t. An individual only decides whether to contribute or not and then gives a pre-specified quantity. The utility from not contributing is normalized to zero. The utilities from contributing are assumed to be

$$\begin{aligned} u_a^t &= \theta_a a + (1-\theta_a)p + s^t - c, && \text{for altruists} \\ u_e^t &= p + s^t - c, && \text{for egoists,} \end{aligned} \tag{1}$$

where $0 < \theta_a \leq 1$. When contributing, both types value the social reward s^t similarly, and they incur the same cost c. Both derive positive utility from money, p. The difference between types lies in the relative importance of money and the altruistic motivation

to contribute, a, where the weight put by egoists on the altruistic incentive is zero.[4]

The social reward for contributing depends on the number of altruistic and egoistic contributors: s^t is a function of N_a^t and N_e^t, $s^t = s(N_a^t, N_e^t) \geq 0$. We assume that the social reward is zero if there are no altruistic contributors and that it increases with their number but it decreases with the number of egoistic contributors. Thus, contributing altruists create positive, and egoists negative externalities for other contributors. Formally, $s(0, N_e^t) = 0$, $\partial s^t / \partial N_a^t > 0$ and $\partial s^t / \partial N_e^t < 0$ for all (N_a^t, N_e^t) such that $s(N_a^t, N_e^t) > 0$. The social reward function lends itself to various interpretations. One possibility is that the source of social reward is belonging to and being recognized as a member of a large group that follows a social norm. The social norm might be "contribute" or "contribute for, at least partly, altruistic reasons". In this case, social reward depends on two factors. One of them is the likelihood of being recognized as a person driven by altruistic motivation, where we assume that people do not know what motivates a contributor. Instead, they take the proportion of altruists in the population of contributors as an estimate of the probability that she is an altruist. The second factor is the number of people who also follow the norm (or only those who do it for altruistic reasons).[5] Some experimental evidence for the assumption can be found in the experimental results of Fehr and Gächter (1999) who find that contributing leads to higher social approval for the contributor if other contributions are high (see also Andreoni and Petrie).

In order to be able to study the dynamics of social change, we assume that in each period one individual is drawn at random to decide whether to contribute. The decisions are made on the basis of the utility from contributing in the previous period: if for a given type it was larger than zero, the selected agent's optimal action is to contribute. Thus, the following dynamics of N_i^t, $i = a, e$, results:

$$\begin{aligned} &\text{If } u_i^t > 0 \quad \text{and} \quad N_i^t < \bar{N}_i, \quad \text{then } N_i^{t+1} = N_i^t + 1 \quad \text{or} \quad N_i^{t+1} = N_i^t. \\ &\text{If } u_i^t < 0 \quad \text{and} \quad N_i^t > 0, \quad \text{then } N_i^{t+1} = N_i^t - 1 \quad \text{or} \quad N_i^{t+1} = N_i^t, \end{aligned} \tag{2}$$

otherwise, N_i^t does not change.

The equilibrium number of egoistic and altruistic contributors is denoted by N_e and N_a, respectively. Given the rules of motion, an equilibrium is reached when $u_i^t = 0$, $i = e, a$, or $u_i^t > 0$ and $N_i = \bar{N}_i$, or $u_i^t < 0$ and $N_i = 0$.

The rules of motion discussed above are based on the assumption that individuals always choose the action that is the best reply to the recent actions of others. As explained in Section 1, we assume that with a small probability ε agents make mistakes and choose the "wrong" action, that is contribute when they should not, or vice versa. This implies that the rules of motion are stochastic: for instance, when the altruists' utility from contributing is

[4] It would be possible, without influencing the results, to introduce a parameter θ_e for egoists, analogous to θ_a, and thus give some intrinsic motivation to egoists as well. However, since we make a sharp distinction between the impact of egoistic and altruistic contributors on the social reward (defined below), it is important that there a clear difference between the two types. The fact that only altruists have intrinsic motivation ensures such a difference, provided that θ_a is large enough.

[5] See Lindbeck et al. (1999) for another model in which social reward increases with the numbers of followers of the norm.

positive, it is most likely that the number of altruistic contributors will increase by one, but there is a small probability that it will actually decrease by one or stay unchanged.

Given the stochastic rules of motion, there is a positive probability of reaching any of the equilibria. If the probability of making mistakes is small enough, however, the short-run evolution of the system will be governed almost surely by the deterministic best-reply dynamics specified above. The equilibrium that arises as a result of this short-run dynamics is termed the *medium-run equilibrium*. We are mainly interested in cases where more than one medium-run equilibrium exist. Which one of these will prevail depends on the initial state. The set of initial states from which the system converges in the medium run to a certain equilibrium with probability one constitutes the *basin of attraction* of that equilibrium.

In the long run, however, it is not the initial state of the system that determines which equilibrium is the most likely to emerge. Rather, the possibility of making a mistake implies that in the long run the system will spend most of the time in the equilibrium that is stochastically stable. When multiple strict equilibria exist, we can use the notions of *radius* and *coradius* of an equilibrium to determine stochastic stability (cf. Ellison, 2000). The procedure is as follows. First, for each equilibrium we find the basin of attraction. Next, for each equilibrium we find the *radius*, defined as the minimum number of errors that are needed to move from that equilibrium out of its basin of attraction. The radius of an equilibrium is then compared to its *coradius*, defined as the maximum over all other states of the minimum number of errors needed to reach the basin of attraction of the equilibrium in question. If there exist equilibria for which the radius is larger than the coradius, all the stochastically stable states will be among them. If there is only one such equilibrium, it will be the unique stochastically stable equilibrium. Intuitively, the radius of an equilibrium describes how difficult it is to get out of it, while its coradius determines how difficult it is to get there from other equilibria. Those equilibria that are more difficult to leave than to reach are stochastically stable.

3. Voluntary contributions

Our analysis starts when no financial reward is provided. Hence, only intrinsic motivation and social reward play a role. Utility functions of both types are given by

$$\begin{aligned} u_a^t &= \theta_a a + s(N_a^t, N_e^t) - c, \\ u_e^t &= s(N_a^t, N_e^t) - c. \end{aligned} \tag{3}$$

We begin with the analysis of the medium-run equilibria and focus first on the deterministic dynamics given by (2). Depending on parameter values and initial states, a variety of equilibria can arise. If $c < \theta_a a$, the intrinsic motivation of an altruist is larger than the cost of contributing, and thus he will contribute no matter what others do. There exists a unique equilibrium in which all altruists contribute, while the number of contributing egoists depends on the size of positive externalities provided by altruists. If, on the other hand, $c > \theta_a a + s(\bar{N}_a, 0)$, so that the costs of contributing are larger than the maximal satisfaction altruists can get, no altruist will ever contribute. Since the egoists' utility from contributing is lower than that of altruists, they will not contribute either. In this case, there is a unique equilibrium with no contributions.

Result 1 describes the more interesting range of parameters where multiple equilibria can arise. The intrinsic motivation alone is not enough to induce an altruist to contribute, but if enough social reward is added, contributing may become worthwhile. Accordingly, three equilibria are possible: one in which no one contributes, one in which some altruists and no egoists contribute, and one in which all altruists contribute. The number of contributing egoists depends on the level of social reward relative to the cost of contributing.

Result 1. If $p = 0$ and $\theta_a a < c < \theta_a a + s(\bar{N}_a, 0)$, three kinds of equilibria exist:

(i) $N_a = N_e = 0$,
(ii) $N_a = N_a^*$ and $N_e = 0$, where N_a^* satisfies $\theta_a a + s(N_a^*, 0) - c = 0$,
(iii) $N_a = \bar{N}_a$, and

(a) $N_e = \bar{N}_e$ if $c < s(\bar{N}_a, \bar{N}_e)$,
(b) $N_e = N_e^*$ if $s(\bar{N}_a, \bar{N}_e) < c < s(\bar{N}_a, 0)$,where N_e^* solves $s(\bar{N}_a, N_e^*) - c = 0$,
(c) $N_e = 0$ If $c > s(\bar{N}_a, 0)$.

Proof. In an equilibrium, either $N_a = 0$, $N_a = \bar{N}_a$ or $0 < N_a < \bar{N}_a$ and $u_a = 0$. We consider these three possibilities in turn. Let u_a, u_e be utility in equilibrium.

(i) If $N_a = 0$, $u_e < u_a < 0$ since $0 < \theta_a a < c$ and we must have that $N_e = 0$. It is easy to see that $N_a = N_e = 0$ is an equilibrium.
(ii) Suppose that $0 < N_a < \bar{N}_a$ and $u_a = 0$. Since $u_e^t < u_a^t$ for any N_a^t and N_e^t, $u_a = 0$ implies that $u_e < 0$. Hence, the only possible equilibrium situation is where $N_e = 0$. Thus, N_a^* must satisfy $u_a = \theta_a a + s(N_a^*, 0) - c = 0$.
(iii) If $N_a = \bar{N}_a$, three values of N_e can arise, depending on the parameter values:

(a) If $c < s(\bar{N}_a, \bar{N}_e)$, the egoists' utility from donating is positive for each $N_e^t \leq \bar{N}_e$. Hence, in equilibrium $N_e = \bar{N}_e$.
(b) If $s(\bar{N}_a, \bar{N}_e) < c < s(\bar{N}_a, 0)$, an egoist gets positive utility from donating if $N_e^t = 0$, but a negative utility if $N_e^t = \bar{N}_e$. It follows that in equilibrium $0 < N_e^* < \bar{N}_e$. Moreover, in equilibrium egoists must be indifferent between donating and not, which yields $u_e = s(\bar{N}_a, N_e^*) - c = 0$.
(c) If $c > s(\bar{N}_a, 0)$, an egoist gets negative utility from donating even if the social reward is maximal. Hence, $N_e = 0$.

It still remains to be shown that when N_e takes these values, $u_a \geq 0$. The three possible cases are $N_e = \bar{N}_e$, $u_e = 0$, or $N_e = 0$. In the first two cases $u_e \geq 0$, which implies $u_a > 0$. When $N_e = 0$, then $u_a = \theta_a a + s(\bar{N}_a, 0) - c$, which is positive by assumption.[6] □

[6] In Result 1 we ignored the integer problem which may appear in case of interior equilibria $0 < N_a < \bar{N}_a$ and $0 < N_e < \bar{N}_e$. Since they must be integer numbers, it may happen that N_a^* and N_e^* satisfying the conditions described in the results do not exist. If there does not exist an integer N_a^* satisfying $u_a = \theta_a a + s(N_a^*, 0) - c = 0$, there does not exist such a number of altruistic contributors at which altruists are indifferent between contributing or not, which implies that the interior equilibrium with $N_a = N_a^*$ does not exist. On the other hand, if there does not exist an integer N_e^* satisfying $u_e = s(\bar{N}_a, N_e^*) - c = 0$, the equilibrium will involve oscillating between N_e equal to the largest integer lower than and the smallest integer larger than N_e^*. This will also hold for all other interior equilibria in the paper.

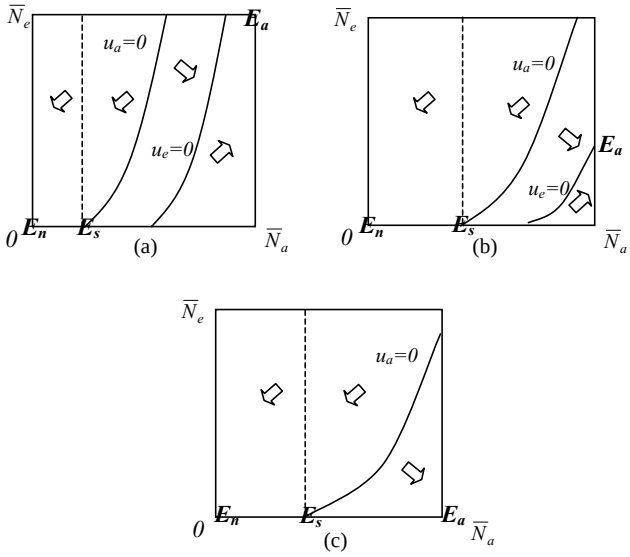

Fig. 1. (a) Medium-run dynamics for $c < s(\bar{N}_a, \bar{N}_e)$; (b) medium-run dynamics for $s(\bar{N}_a, \bar{N}_e) < c < s(\bar{N}_a, 0)$; (c) medium-run dynamics for $c > s(\bar{N}_a, 0)$.

In case of multiple equilibria, the medium run outcome depends on the initial state. This is illustrated in Fig. 1a–c for three different ranges of parameter values. The horizontal and vertical axes show the number of altruistic and egoistic contributors, respectively, ranging from 0 to their total numbers in the whole population. E_a, E_s and E_n denote equilibria with all, some and no altruists contributing, respectively. The indifference curves, $u_a = 0$ and $u_e = 0$, show combinations of N_a^t and N_e^t at which altruists and egoists, respectively, are indifferent between contributing and not contributing. As social reward increases in N_a^t and decreases in N_e^t, the indifference curves are upward sloping. Note also that $s(0, N_e^t) = 0$ implies that if an indifference curve shows up in the figure, it must cross the bottom borderline, $N_e^t = 0$. The exact shape of the indifference curves depends on the shape of the social reward function.

The altruists' utility from contributing is positive to the right of their indifference curve, and negative to the left of it. Hence, according to the rules of motion, N_a^t is increasing to the right and decreasing to the left of the curve. This follows from the positive externalities generated by altruists that cause a critical mass effect: when the number of altruists exceeds a certain critical mass, all other altruists are attracted. On the other hand, the utility of egoists is positive below their indifference curve, and negative above. Hence, N_e^t is typically converging to an interior equilibrium value. This is the negative externality at work. The dynamics of N_a^t and N_e^t is illustrated by arrows. As $\theta_a a < c < \theta_a a + s(\bar{N}_a, 0)$ in all three figures, the altruists' indifference curve shows up, and three equilibria exist. In Fig. 1c, $u_e < 0$ for all combinations of N_a^t and N_e^t, which is why the egoists' indifference curve does not show up in that figure. One can easily see that the equilibrium with only some altruists contributing is unstable.

When the society is initially to the right of the altruists' indifference curve, it will move in the medium run towards the equilibrium with all altruists contributing. Hence, this area is the basin of attraction of that equilibrium. Similarly, the area to the left of the dashed line $N_a^t = N_a^*$ is the basin of attraction of the equilibrium with no contributors. For these initial states the number of altruistic contributors is too small to attract other altruists, so eventually all altruists will stop contributing. Finally, when the society is initially in one of the remaining states, either E_a or E_n can arise with positive probability. Hence, in the medium run with deterministic dynamics, the outcome depends on the initial state.

In the long run, when people make occasional mistakes against their interest, any of the equilibria can arise with a positive probability, independently of the initial state. In particular, the equilibrium with contributors can arise and persist in the long run even if the initial level of contributions was zero. For this to happen it is sufficient that the equilibrium with contributors is stochastically stable. Applying the radius-coradius method described in Section 2 easily shows that the equilibrium with all altruists contributing is stochastically stable if the distance (in terms of number of errors) from E_a to the altruists' indifference curve is larger than the distance from E_n to that curve. That is most likely to occur if $\theta_a a$ is relatively large, which implies a low value of N_a in E_s, and if the value of N_e in E_a is small. In other words, voluntary contributions are most likely to emerge if the altruistic motivation of altruists is large and if not too many egoists are induced to contribute by positive externalities created by altruists.

In the rest of the paper we will focus on the case where multiple medium-run equilibria exist and the equilibrium in which all altruists contribute is stochastically stable. In this case a social norm to contribute will arise in the long run, and it may be crowded out by a financial reward. For other parameter values the outcomes are more obvious and thus less interesting.

4. Introducing a financial reward

In this section, we study the effect of introducing a financial reward into the situation analyzed in the previous section: from now on we assume that $p > 0$. As we consider the case where the equilibrium with contributors is stochastically stable, that will be the initial situation at the moment a financial reward is introduced.

First, we consider the medium-run dynamics, where all agents behave optimally. Results 2 and 3 describe the new medium-run equilibria, for two ranges of parameter values. Note that since now $p > 0$, the utility functions are as in Eq. (1).

Result 2. Suppose that $\theta_a a < c < a + s(\bar{N}_a, \bar{N}_e)$. When $p > 0$ is introduced, in the new medium-run equilibrium $N_a = \bar{N}_a$, and

(i) $N_e = \bar{N}_e$ if $p > c - s(\bar{N}_a, \bar{N}_e)$,
(ii) $N_e = N_e^{**}$ if $c - s(\bar{N}_a, 0) < p < c - s(\bar{N}_a, \bar{N}_e)$,
where N_e^{**} satisfies $p + s(\bar{N}_a, N_e^{**}) - c = 0$,
(iii) $N_e = 0$ if $p < c - s(\bar{N}_a, 0)$.

Proof. We first assume that in the equilibrium $N_a = \bar{N}_a$, and we show that this implies N_e as described above. Next, we show that if $N_a = \bar{N}_a$ and N_e is as above, $u_a > 0$, and thus $N_a = \bar{N}_a$. Suppose that $N_a = \bar{N}_a$. We have to consider three cases.

(i) If $p > c - s(\bar{N}_a, \bar{N}_e)$ and $N_a = \bar{N}_a$, egoists get positive utility for any N_e^t. Hence, $N_e = \bar{N}_e$.

(ii) If $c + s(\bar{N}_a, 0) < p < c + s(\bar{N}_a, \bar{N}_e)$ and $N_a = \bar{N}_a$, egoists get negative utility if $N_e = \bar{N}_e$ but positive utility if $N_e = 0$. Hence, in the equilibrium $0 < N_e^{**} < \bar{N}_e$ and $u_e = p + (\bar{N}_a, N_e^{**}) - c = 0$.

(iii) If $p < c - s(\bar{N}_a, 0)$, egoists have a negative utility for any N_e. Hence, $N_e = 0$. We still have to show that in the new equilibrium $u_a > 0$. Note first that if $p > a$, $c < a + s(\bar{N}_a, \bar{N}_e)$ implies $u_a^t = \theta_a a + (1 - \theta_a)p + s(\bar{N}_a, N_e^t) - c > 0$ for each N_e^t. Suppose now that $p < a$. Then, $p < \theta_a a + (1 - \theta_a)p$, which implies that if in the equilibrium $u_e \geq 0$, then also $u_a > 0$. If in the equilibrium $u_e < 0$, $N_e = 0$ and $u_a = \theta_a a + (1 - \theta_a)p + s(\bar{N}_a, 0) - c > \theta_a a + s(\bar{N}_a, 0) - c > 0$ by assumption. □

Result 2 shows that as $c < a + s(\bar{N}_a, \bar{N}_e)$, altruists can never be discouraged from contributing. This is because starting from the initial situation, egoists will only donate when $c < p + s(\bar{N}_a, 0)$. If p is small, the number of egoists attracted to contributing is small and so are the negative externalities they generate, but if p is large, it provides a sufficient incentive to contribute also for altruists. Hence, the introduction of a financial reward either leads to an increase in the amount of contributions, or it does not have an impact. On the other hand, the social reward may decrease so much that it becomes insufficient (if not combined with a monetary reward) to encourage altruists to contribute. We will see in Section 5 that if such weakening of the social reward occurs and the financial reward is withdrawn, contributions may fall to zero.

Fig. 2 shows an example of the dynamics process after a price is introduced.

In the figure, $u_a^0 = 0$ and $u_e^0 = 0$ denote the old indifference curves, E_a^0 is the old equilibrium with all altruists contributing, and E_a, E_s and E_n are the new equilibria. The parameters are chosen such that in E_a^0, $0 < N_e^0 < \bar{N}_e$, while $N_e = \bar{N}_e$ in E_a. The black arrow shows the path from E_a^0 to E_a. Introducing a price $p > c - s(\bar{N}_a, 0)$ shifts both indifference curves upwards. The number of egoistic contributors increases, but since after this change $u_a > 0$, the number of altruistic contributors remains $\bar{N}_a$ and E_a is the new equilibrium.

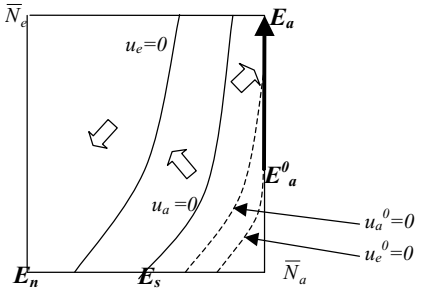

Fig. 2. Medium-run dynamics for $\theta_a a < c < a + s(\bar{N}_a, \bar{N}_e)$ and $p > a$.

In the next Result we deal with the case in which a financial reward of an intermediate size may decrease the amount of contributions in the medium run. Let N_e^{**} be defined as in Result 2.

Result 3. Suppose that $a + s(\bar{N}_a, \bar{N}_e) < c < \theta_a a + s(\bar{N}_a, 0)$. When a $p > 0$ is introduced, the new medium-run equilibrium is

(i) $N_a = \bar{N}_a$ if $p < a$, and

  (a) $N_e = N_e^{**}$ if $p > c - s(\bar{N}_a, 0)$,
  (b) $N_e = 0$ if $p < c - s(\bar{N}_a, 0)$.

(ii) $N_a = 0$ if $a < p < [c - \theta_a a - s(\bar{N}_a, \bar{N}_e)]/(1 - \theta_a)$, and

  (a) $N_e = 0$ if $p < c$,
  (b) $N_e = \bar{N}_e$ if $p > c$.

(iii) $N_a = \bar{N}_a$ and $N_e = \bar{N}_e$ if $p > [c - \theta_a a - s(\bar{N}_a, \bar{N}_e)]/(1 - \theta_a)$.

Proof.

(i) Suppose $p < a$. This implies that $u_a^t > u_e^t$ for any N_a^t and N_e^t. This also implies that the egoists' utility is negative when $N_e = \bar{N}_e$, since then $u_e < u_a = \theta_a a + (1-\theta_a)p + s(\bar{N}_a, \bar{N}_e) - c < a + s(\bar{N}_a, \bar{N}_e) - c < 0$. Thus, in the equilibrium either $u_e = 0$, or $N_e = 0$. In case (a), $p > c - s(\bar{N}_a, 0)$ implies that the egoists' utility is positive when $N_e = 0$, since $u_e = p + s(\bar{N}_a, 0) - c > 0$. Hence, in the equilibrium $u_e = 0$ and $N_e = N_e^{**}$ which satisfies $u_e = p + s(\bar{N}_a, N_e^{**}) - c = 0$. At this value of N_e, altruists will still prefer to contribute as $u_a > u_e = 0$. Consider now case (b). If $p < c - s(\bar{N}_a, 0)$, egoists' utility from contributing is negative even if $N_e = 0$, and hence no egoist will contribute. The utility of altruists is $u_a = \theta_a a + (1-\theta_a)p + s(\bar{N}_a, 0) - c > \theta_a a + s(\bar{N}_a, 0) - c > 0$, and thus $N_a = \bar{N}_a$.

(ii) Suppose to the contrary that $N_a > 0$. This implies that $u_a \geq 0$, but $p > a$ implies $u_e^t > u_a^t$ for any N_a^t and N_e^t, which in turn implies that $u_e > 0$ and $N_e = \bar{N}_e$. But then $u_a \leq \theta_a a + (1 - \theta_a)p + s(\bar{N}_a, \bar{N}_e) - c < 0$, which contradicts $N_a > 0$. Hence, $N_a = 0$. Then, $u_e = p - c$, and thus $N_e = \bar{N}_e$ if $p > c$, and $N_e = 0$ if $p < c$.

(iii) $p > [c - (\bar{N}_a, \bar{N}_e) - \theta_a a]/(1 - \theta_a)$ implies that when $N_a = \bar{N}_a$, altruists' utility is positive for every N_e. Hence, $N_a = \bar{N}_a$. From $p > a$ it follows that $u_e^t > u_a^t > 0$ for any N_a^t and N_e^t, which implies $N_e = \bar{N}_e$. □

In the situation described in Result 3, a large number of egoistic contributors may discourage altruists from contributing. The impact of a price on total contributions is non-monotonic. If the price is low, no or few egoists are attracted, negative externalities are small, and altruists continue to contribute. Thus, the effect on the amount of contributions is either none, or positive and small. If the price is high, the financial reward outweighs (also for altruists) the decrease in the social reward due to negative externalities created by egoistic contributors, and the amount of contributions increases or stays unchanged. The crowding out effect occurs when the price takes intermediate values, high enough to attract many egoists, but too low to compensate the loss of the social reward to altruists.

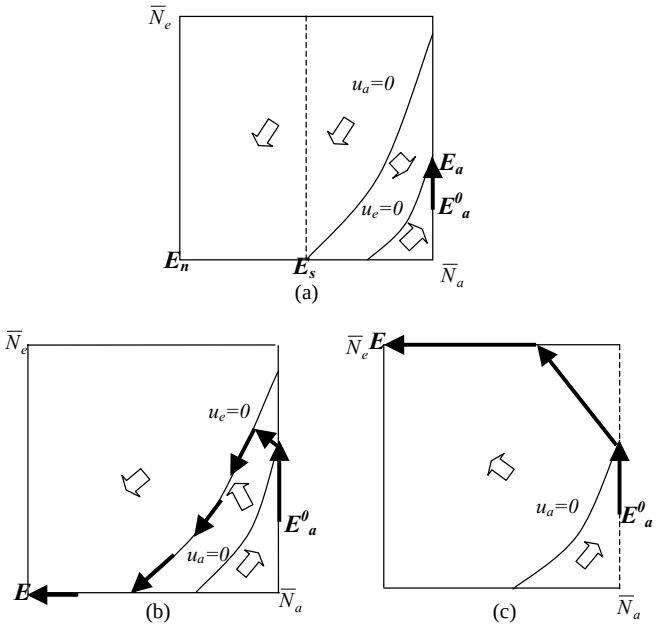

Fig. 3. (a) Medium-run dynamics: case (i) (b), $c - s(\bar{N}_a, 0) < p < a$; (b) case (ii) (a) $a < p < [c - s(\bar{N}_a, \bar{N}_e) - \theta_a a]/(1 - \theta_a)$ and $p < c$; (c) case (ii) (b) $c < p < [c - s(\bar{N}_a, \bar{N}_e) - \theta_a a]/(1 - \theta_a)$.

The altruists stop contributing, the norm disappears, and the only remaining motivation for egoists is the financial reward. If it does not exceed the cost of contributing, egoists stop contributing as well, and the total contributions fall to zero. This is the worst possible situation: the social norm is crowded out, and the financial reward by itself is not large enough to compensate for the cost. If the price is higher than the cost, egoists contribute in the new equilibrium. In comparison with the old equilibrium, the nature of a typical contributor has changed, however: before he was likely to be an altruist, now he surely is an egoist. Total contributions may decrease or increase, depending on the total numbers of altruists and egoists in the population.

The medium-run dynamics of the system in cases (i) and (ii) of Result 3 are illustrated in Fig. 3a–c. Note that when $p > a$, there is only one new equilibrium, denoted by E. Again, the black arrows show the transition from the old to the new equilibrium. Introducing a financial reward shifts both indifference curves upwards. If $p < a$ (as in Fig. 3a), the egoists' indifference curve stays below that of the altruists. This means that the price does not attract enough egoists to make altruistic contributors change their behavior. If $a < p < [c - \theta_a a - s(\bar{N}_a, \bar{N}_e)]/(1 - \theta_a)$ and $p < c$ (as in Fig. 3b), the egoists' indifference curve lies above that of altruists. In that case, crowding out occurs: when a price is introduced, initially the number of altruistic contributors is constant, and the number of egoistic contributors is increasing. When the system reaches the indifference curve of altruists, the number of altruistic contributors starts decreasing, but the number of egoists keeps increasing until the system reaches the egoists' indifference curve. From that moment onwards, both N_a^t

and N_e^t gradually fall to zero. If $c < p < [c - s(\bar{N}_a, \bar{N}_e) - \theta_a a]/(1 - \theta_a)$ (as in Fig. 3c), the egoists utility from supplying is always positive and thus N_e^t increases to $\bar{N}_e$, while the number of altruistic contributors decreases and eventually falls to zero. The last case, $p > [c - s(\bar{N}_a, \bar{N}_e) - \theta_a a]/(1 - \theta_a)$, in which the price is high enough so that the utility of all individuals is positive and everyone contributes, is not illustrated with a figure.

Let us now turn to the long-run analysis. If $p > a$, there is a unique equilibrium, either with $N_a = 0$ or with $N_a = \bar{N}_a$, which must then arise also in the long run. For a lower p both E_a and E_n exist. Just as in Section 3, the stochastic stability of E_a is determined by the comparison of the smallest distance (in terms of the number of errors) from each equilibrium to the altruists' indifference curve. As compared to the situation before the financial reward was introduced, the new equilibrium with contributors may lose or gain the property of stochastic stability. Given N_e, the financial reward increases utility of altruists from contributing for any (N_a^t, N_e^t), which shifts their indifference curve leftwards and has a positive impact on the stochastic stability of the equilibrium with altruistic contributors. However, the price may increase N_e as well, decreasing the social reward and the utility of altruists and therefore having a negative impact on the stochastic stability of E_a. The eventual net effect of a price depends on the parameter values and the shape of the reward function. For example, the equilibrium with contributors becomes less stable after the introduction of a price if $\theta_a = 1$, and originally $0 < N_e < \bar{N}_e$. In this case, a financial reward does not have any direct effect on the altruists' utility, only a negative indirect effect through the increased number of egoists. Since all the relevant distances are continuous in θ_a, the same conclusion holds for θ_a close to 1. Thus, it is possible that due to the introduction of a price the new equilibrium becomes stochastically unstable.

5. Withdrawing the financial reward

In the previous section we saw that introducing a monetary incentive may have an adverse effect on the level of contributions. When the authorities realize that the social norm has disappeared, they may decide to abandon the payment in order to restore the previous situation. In this section, we show that this may fail to work. Withdrawing the financial reward shifts the indifference curves back to the old positions. The same two medium-run equilibria exist as before the payment was introduced. Thus, when the equilibrium with contributors is stochastically stable, it will reappear in the long run. However, in the medium run, the society may not return to the equilibrium with altruistic contributors, but instead move towards the equilibrium without any contributions. This happens when at the moment of withdrawing the payment the system is located in the basin of attraction of E_n, the equilibrium without contributions.

It is easy to see that this will be the case if altruists did not contribute when $p > 0$. In this case the social reward for contributing had been crowded out, and when the financial reward is also removed, there is no immediate reason to contribute. Thus, precisely in those cases when the financial reward has had a negative impact and authorities might be most tempted to remove it, its withdrawal does not make the situation any better, and it may even make it worse. Result 4 states this formally. In all results and proofs of this section, let N_a^0

and N_e^0 denote the number of altruistic and egoistic contributors in the equilibrium with a financial reward $p_0 > 0$, and u_a^0 and u_e^0 their utilities in that equilibrium.

Result 4. Suppose that $\theta_a a < c < \theta_a a + s(\bar{N}_a, 0)$, initially $p_0 > 0$ and $N_a^0 = 0$. When the financial reward is withdrawn, the new medium-run equilibrium has $N_a = 0$ and $N_e = 0$.

Proof. If $N_a^0 = 0$, $u_a^0 = \theta_a a + (1 - \theta_a)p_0 + s(0, N_e^t) = \theta_a a + (1 - \theta_a)p_0 < 0$. Thus, $u_a^t < 0$ for all $p < p_0$ and all N_e^t. It follows that in the new equilibrium $u_a < 0$ and $N_a = 0$. When $p = 0$ and $N_a^0 = 0$, $N_e = 0$ follows. □

On the other hand, if the social norm had not been crowded out completely, the effect of withdrawing the financial reward is ambiguous. If the price is low so that not many egoists are attracted, or if the social reward is still strong despite many egoistic contributors, the social reward stays large enough to provide sufficient incentives to contribute for altruists. Then, withdrawing the financial reward does not lower the altruists' motivation enough to make them stop contributing. In other cases, a monetary incentive is necessary to motivate the altruists.

Let us now describe these results more formally. When a gradual price adjustment is allowed, the eventual outcome may depend on the price adjustment process. Denote a sequence of prices converging to zero by $p_0 > p_1 > \cdots > p_K = 0$, where p_0 is the original price. When the price is withdrawn immediately, $K = 1$. To make the analysis tractable, we assume that an intermediate price is introduced only after the adjustment process following the previous price change is completed and an intermediate equilibrium has been reached. Denote these intermediate equilibria by (N_a^1, N_e^1), (N_a^2, N_e^2), ..., etc.

Result 4 still holds when the price is reduced gradually. That is, for every sequence of prices converging to zero, the equilibrium (0,0) eventually arises. Result 5 below describes the equilibria that will arise in the medium run after the financial reward has been gradually removed (in case two equilibria are possible), thus $\theta_a a < c < \theta_a a + s(\bar{N}_a, 0)$, and when in the original equilibrium $N_a^0 = \bar{N}_a$. The result states that for certain parameter values a gradual reduction of the financial reward restores the equilibrium from before its introduction.

Result 5. Suppose that $\theta_a a < c < \theta_a a + s(\bar{N}_a, 0)$, $p_0 > 0$ and $N_a^0 = \bar{N}_a$. There exists a sequence of prices $p_0 > p_1 > \cdots > p_K = 0$ such that in the eventual medium-run equilibrium $N_a^K = \bar{N}_a$ with probability one if and only if $p_0 < a$ or $c < a + s(\bar{N}_a, \bar{N}_e)$.

Proof. It follows from the proof of Result 4 that if in an intermediate equilibrium $N_a^k = 0$, it is also the case in all subsequent equilibria. Thus, if in the eventual equilibrium $N_a^K = \bar{N}_a$, it must be the case that $N_a^k = \bar{N}_a$ in all intermediate equilibria. Of course, if $c < s(\bar{N}_a, \bar{N}_e)$, then one can withdraw the price immediately and have $N_a^K = N_e^K = \bar{N}_a$ in a trivial way. We now consider therefore the case where $c > s(\bar{N}_a, \bar{N}_e)$.

If $N_a^K = \bar{N}_a$ is to arise with probability one, every pair of intermediate prices p_{k-1} and p_k must be such that the intermediate equilibrium $(\bar{N}_a, N_e^{k-1})$ lies in the basin of attraction of the intermediate equilibrium $(\bar{N}_a, N_e^k)$. Moreover, we can safely restrict attention to price changes that are such that $N_e^k \neq N_e^{k-1}$. It follows that we have to consider a price sequence

$\{p_k\}$ such that for every $(\bar{N}_a, N_e^{k-1})$, the altruists' utility from contributing is positive, i.e. $u_a^{k,k-1} \equiv \theta_a a + (1-\theta_a)p_k + s(\bar{N}_a, N_e^{k-1}) - c > 0$ and the egoists' utility from contributing is negative, $u_e^{k,k-1} \equiv p_k + s(\bar{N}_a, N_e^{k-1}) - c < 0$. The question is under what conditions such a price sequence exists.

There are two cases to consider. In the first case, $N_e^0 < \bar{N}_e$, implying that $p_0 < a$. Then, $u_a^k > u_e^k$ for every $p_k \leq p_{k-1}$. Moreover, for every p_k, N_e^k adjust in such a way that $u_e^k = 0$. Thus, there exists a price path such that in every intermediate and in the eventual equilibrium $N_a^K = \bar{N}_a$ with probability one.

The second case is $N_e^0 = \bar{N}_e$. It is clear that we have to choose $p_1 < a$. Moreover, given that $c > s(\bar{N}_a, \bar{N}_e)$ it easy to see that a price p_1, with $0 < p_1 < a$ that satisfies both conditions mentioned above, exists if and only if $a + s(\bar{N}_a, \bar{N}_e) - c > 0$. Choosing such a price results in $N_e^1 < \bar{N}_e$ and the rest of this case is similar to the case where $N_e^0 < \bar{N}_e$. □

It follows from Result 5 that in certain cases, for instance when $\theta_a a < p_0 < a$, by withdrawing the price gradually rather than immediately, authorities can make sure that the social norm does not disappear and the amount of contributions remains positive after the price is withdrawn completely. For this to happen, they have to choose a price sequence such that the society always stays in the basin of attraction of the equilibrium with contributors. In other words, they should always choose a price that is low enough to discourage some egoists to contribute, but also high enough to keep the altruists contributing. After the number of egoistic contributors has decreased and the social reward has increased, the price becomes less important as a source of motivation for altruists and can be lowered further.

Thus, a gradual reduction of the price may in some cases help to restore a social norm that has not been crowded out, but weakened by the financial reward. However, this is not always true. Recall the crowding out result from Section 4. In Result 3 we stated that if $c > a + s(\bar{N}_a, \bar{N}_e)$, introducing a price satisfying $a < p < [c - \theta_a a - s(\bar{N}_a, \bar{N}_e)]/(1-\theta_a)$ leads to an equilibrium in which $N_a = 0$. Since for this price this is the only equilibrium, this outcome will always arise whether the price is raised to that level from zero or lowered from a high level.

6. Crowding in

Suppose that due to the introduction of a financial reward, or its subsequent withdrawal, the norm disappears. We have shown in Section 3 that the norm can arise in the long run again if $p = 0$, provided that the equilibrium with contributors is stochastically stable. However, the authorities do not always have to wait for the social reward to reappear spontaneously as they could try to crowd it in (bring it into existence) with a financial reward. The possibility of such crowding in only occurs if altruists care about money. In this case, there exists a price high enough to encourage them to contribute that brings about a positive social reward. With this reward in place, it may be possible to reduce the price to zero without discouraging altruists from contributing. For this to happen, it must be possible to decrease the price enough to discourage egoists from contributing without at the same time discouraging altruists.

Let us now analyze crowding in in some more detail. The price that encourages altruists to contribute even in the absence of the social reward must be such that $u_a > 0$, or,

$$p > \frac{c - \theta_a a}{1 - \theta_a}. \tag{4}$$

When this price is introduced, $N_a = \bar{N}_a$, implying that the social reward is positive. Thus, even if the price is somewhat reduced, the altruists will keep contributing. Specifically, it is always possible to keep $N_a = \bar{N}_a$ while reducing the price to the level

$$p \geq \frac{c - \theta_a a - s(\bar{N}_a, \bar{N}_e)}{1 - \theta_a} \tag{5}$$

In some cases p can be lowered even further. We are most interested in the possibility of crowding in the social reward sufficiently to make the financial reward redundant to secure contributions. The consequences of withdrawing the price for different original price levels, initial states and parameter values have already been described in detail in Results 4 and 5 of Section 5. Here, we only recuperate those outcomes that are relevant for crowding in. Assume, as before, that a new price is always introduced after the adjustment to the intermediate equilibrium under the previous price has taken place. Note that since after the introduction of a price $N_a = \bar{N}_a$, there is always a positive probability that crowding in will be successful. Result 6 describes the conditions under which crowding in can occur, under deterministic dynamics, with certainty.

Result 6. Suppose that $\theta_a a < c < \theta_a a + s(\bar{N}_a, 0)$ and initially $p_0 = 0$ and $N_a^0 = N_e^0 = 0$. Then, if and only if $c < a + s(\bar{N}_a, \bar{N}_e)$ there exists a price sequence $p_1 > p_2 > \cdots > p_K = 0$, $p_1 > (c - \theta_a a)/(1 - \theta_a)$, such that in the eventual medium-run equilibrium $N_a = \bar{N}_a$ with probability one.

Proof. This follows straightforwardly from (4) and Result 5. If $p_1 > 0$ and $N_a^1 = \bar{N}_a$, such a price sequence exists if and only if $p_1 < a$ or $c < a + s(\bar{N}_a, \bar{N}_e)$. Since $N_a^1 = \bar{N}_a$ requires $p_1 > (c - \theta_a a)/(1 - \theta_a)$, $p_1 < a$ is only possible if $c < a < a + s(\bar{N}_a, \bar{N}_e)$. □

Note that the social reward can be crowded in with certainty only for those parameter values for which it cannot be crowded out by the introduction of a financial reward (see Result 2). On the other hand, in those cases in which crowding out can take place, crowding in is uncertain. Introducing a high price and then reducing it to zero may lead to a creation of the social reward, but the outcome of the process is uncertain. Thus, precisely in those cases in which crowding out has occurred, the possibilities of using financial incentives to (re)create the social reward are limited.

7. Discussion and conclusions

With the help of a simple evolutionary model, we have given an interpretation to a number of empirical findings: a social norm to contribute may disappear or weaken after a financial reward is introduced, and if it has been destroyed, it can take a long time before

it re-emerges. We have also shown how the norm could have arisen in the first place in the absence of a financial reward, namely as the result of a long history with a voluntary system of contributing. Our results point to potential dangers hidden in the use of insufficient financial incentives in situations in which social norms play an important role. Even if the social norm is not crowded out immediately, it may become more fragile due to the use of financial rewards and, therefore, disappear in the long run. It may also become insufficient to secure contributions in the absence of a payment.

Does it mean that existing financial rewards should be removed? In the short and medium run, it will often make the situation even worse. More specifically, if a social norm has already been crowded out, removing the payment leads to the breakdown of contributions in the medium run. Hence, if waiting for the norm to reappear is not a feasible option, it makes more sense to keep the reward in place. If the social norm has not (yet) disappeared, the situation is more ambiguous. If the price has made the social reward very low, withdrawing it may result in a collapse of contributions. On the other hand, in some cases removing the price may prevent the norm from disappearing in the long run.

Finally, for the situation where no social reward exists (possibly because it has been crowded out) we have examined the possibility of creating, with the help of a price, a social reward strong enough to sustain contributions even after the financial incentive is removed. It turns out that such crowding in is possible, although it is least effective when most needed and it often requires that authorities have very detailed information about utilities of individuals, which may not be available in practice. In addition, a successful crowding in may require a very high initial price. Since in many cases the actual amounts offered for contributing are relatively low, a sufficiently high price may not always be feasible.

The assumption that egoists care about the social reward is not crucial for either crowding out or crowding in: it makes the results richer by allowing egoists to contribute only to gain social reward, and it reduces the differences between agents, but both effects rely mainly on discouraging or encouraging the participation of altruists. In addition, $a > 0$ is not crucial; hence, it is not necessary that altruists are really "altruistic", although it is important that they derive less utility from money than egoists.

Acknowledgements

Earlier versions of this paper have been presented at: The Spring Meeting of Young Economists, Copenhagen, 2001; The 16th Annual Congress of the European Economic Association, Lausanne; NAKE Research Day 2001 and an Economic Theory seminar, Erasmus University, 2002. In addition to the members of the audiences, we thank Robert Dur, Vladimir Karamychev, Bauke Visser and Jurjen Kamphorst for their useful comments.

References

Andreoni, J., Petrie, R., 2000. Social motives to giving: can these explain fund raising institutions? mimeo.
Arrow, K.J., 1972. Gifts and exchanges. Philosophy and Public Affairs 1, 343–362.
Bar-Gill, O., Fershtman, C., 2001. The limit of public policy: endogenous preferences. mimeo.

Bester, H., Güth, W., 1998. Is altruism evolutionarily stable? Journal of Economic Behavior and Organization 34, 193–209.
Deci, E., 1999. Meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. Psychological Bulletin 125, 627–668.
Dickinson, A., 1989. The detrimental effects of extrinsic reinforcement on 'intrinsic motivation'. The Behavior Analyst 12, 1–15.
Ellison, G., 2000. Basins of attraction, long-run stochastic stability, and the speed of step-by-step evolution. Review of Economic Studies 67, 17–45.
Fehr, E., Gächter, S., 1999. Collective action as a social exchange. Journal of Economic Behavior & Organization 39, 341–369.
Fehr, E., Gächter, S., 2000. Do incentive contracts crowd out voluntary cooperation? Working Paper No. 34. Insititute for Empirical Research in Economics, University of Zurich.
Frank, R.H., 1987. If homo economicus could choose his own utility function, would he choose one with a conscience? American Economic Review 77, 593–604.
Frey, B., 1997. Not just for the money. Edward Elgar Publishing, Cheltenham, UK, Brookfield, US.
Frey, B.S., Götte, L., 1999. Does pay motivate volunteers? Working Paper No. 7. Insititute for Empirical Research in Economics, University of Zurich
Frey, B., Jegen, R., 2001. Motivation crowding theory: a survey of empirical evidence. Journal of Economic Surveys 15, 589–611.
Glazer, A., Konrad, K., 1996. A signalling explanation for charity. American Economic Review 86, 1019–1028.
Gneezy, U., Rustichini, A., 2000a. A fine is a price. Journal of Legal Studies 29, 1–17.
Gneezy, U., Rustichini, A., 2000b. Pay enough—or don't pay at all. Quarterly Journal of Economics 115, 791–810.
Güth, W., Kliemt, H., 2000. Evolutionarily stable co-operative commitments. Theory and Decision 49, 197–221.
Harbaugh, W.T., 1998a. What do contributions buy? A model of philantropy based on prestige and warm glow. Journal of Public Economics 67, 269–284.
Harbaugh, W.T., 1998b. The prestige motive for making charitable transfers. American Economic Review 88, 277–282.
Lindbeck, A., Nyberg, S., Weibull, J., 1999. Social norms and economic incentives in the welfare state. The Quarterly Journal of Economics 114, 1–35.
Stewart, H., 1992. Rationality and the market for human blood. Journal of Economic Behavior and Organization 19, 125–143.
Titmuss, R.M., 1970. The Gift Relationship: From Human Blood to Social Policy. George Allen and Unwin, London.
Weibull, J.W., 1995. Evolutionary Game Theory. MIT Press, Cambridge, London.
Young, P., 1998. Individual Strategy and Social Structure. Princeton University Press, Princeton.