FAERE

French Association
of Environmental and Resource Economists

Working papers

An instrument that could turn crowding-out into crowding-in

Antoine Beretti – Charles Figuières –
Gilles Grolleau

WP 2014.04

Suggested citation:

Beretti A., Figuières C. and Grolleau G. (2014). An instrument that could turn crowding-out into crowding-in. FAERE Working Paper, 2014.04.

ISSN number:

http://faere.fr/working-papers/

An instrument that could turn crowding-out into crowding-in

Antoine Beretti[*], Charles Figuières[†] and Gilles Grolleau[‡]

April 8, 2014

Abstract

Using a simple decision-theoretic approach, we formalize how agents with different kinds of intrinsic motivations react to the introduction of monetary incentives. We contend that empirical results supporting the existence of a crowding-out effect in various contexts hide a more complex reality. We also propose a new policy instrument which taps into agents' heterogeneity regarding intrinsic motivations in order to turn a situation subject to crowding-out into a crowding-in outcome. This instrument uses a self-selection mechanism to match adequate monetary incentives with individuals' types regarding intrinsic motivations.

Keywords: crowding-out, heterogeneity, moral motivation, environmental regulation.

JEL Classification: D03, D64, H23, Q58.

[*]LAMETA - SupAgro.

[†]UMR LAMETA, INRA, place Pierre Viala, Bât. 26, 34060 Montpellier Cedex 1, France. Email: Charles.Figuieres@supagro.inra.fr

[‡]LAMETA - SupAgro.

1 Introduction

Monetary incentives matter and are a powerful instrument to change behaviour. They are a part of the story but not all the story. Indeed, the crowding-out effect, which has recently received considerable attention in economic literature, stresses that the preferred leverage of economists, monetary incentives and disincentives can backfire and lead to inferior outcomes. Such a counter-productive effect could be due to an interaction - still poorly understood - between intrinsic motivation and extrinsic (dis)incentives introduced by the monetary instrument. The issue has been investigated, theoretically (Bénabou and Tirole, 2003) and empirically (see Bowles, 2008 for a recent review). It has been proven to be relevant in a wide variety of contexts such as blood donation (Titmuss, 1970; Mellstrom and Johannesson, 2008, Goette, Stutzer & Frey, 2010), acceptance of a polluting infrastructure (Frey and Oberholzer-Gee, 1997) or to address late coming parents in day-care centers (Gneezy and Rustichini, 2000) or bed-blocking in hospitals (Holmås et al., 2010).

Most papers consider that all agents behave similarly when facing monetary incentives. Nevertheless, it is more realistic to assume that people are heterogeneous and have various intrinsic motivations according to the considered domain (Bénabou and Tirole, 2006; Beretti et al., 2013). This heterogeneity is probably the source of many difficulties encountered in defining the concept of intrinsic motivation, and controversies about its usefulness (Reiss, 2006; Bruno, 2012). Psychologists frequently consider that intrinsic motivations are those that arise from within – doing something because you want to – while extrinsic motivations mean people are seeking a reward, such as money or a trophy at a sporting event. Intrinsic motivation is that which is pleasurable *per se*, while extrinsic motivation is not. Put differently, one could resort to a means-end logic in order to determine whether a motivation is intrinsic or extrinsic: intrinsic motivation is doing what we want, whereas extrinsic motivation is doing something to get something else. Recently, Bolle and Otto (2010) proposed to define extrinsic motivations as added motivations that interfere and disturb a well-defined situation. This well-defined situation constitutes the reference where intrinsic motivations play their role. Their definition along with others are collected in Table 1 below. To measure intrinsic motivation, various approaches have been used such as self-reported measures of engagement and interest in the activity, observation of free-choice engagement in the activity when no rewards or other extrinsic motivators are present. Moreover, there are some studies investigating the possible neural basis of intrinsic motivation (e.g., Lee et al, 2012). They report differences in neural

activation between intrinsic and extrinsic motivations.

Table 1. Intrinsic motivation: definitions

Definitions	Authors
Intrinsically motivated behaviors are 'behaviors in which a person engages in to provide himself with a sense of competence and self-determination'	Deci, 1975
Intrinsic motivation is a 'measure of task engagement in a situation in which salient extrinsic contingencies had been deliberately minimized'	Lepper, 1978
Intrinsic motivation 'means enjoying what one does for its own sake'	Kohm, 1993
'To be intrinsically motivated means to engage in an activity because the activity itself is interesting and enjoyable'	Deci et al., 2008
'Extrinsic motivation can be defined 'as stemming from a change in the environment'	Bolle & Otto, 2010

We argue that any agent shelters a unique combination of various intrinsic motivations, and that this individual heterogeneity across agents is not without consequences on the effectiveness of public interventions. In particular, its overall impact in a whole population supporting or not the existence of a crowding-out effect can overlook the fact that various subgroups of this whole population react differently when facing the introduction of monetary incentives. Indeed, moderate monetary incentives can both motivate people who were originally not intrinsically motivated and harm the intrinsic motivations of people who were originally intrinsically motivated. Several studies reveal that incentive-based programs do not produce identical reactions across individuals (Gneezy et al., 2011) and psychologists recognize that different people are motivated in different ways (Reiss, 2006; Lindenberg, 2001). Natural candidates for this heterogeneity can be related to contextual parameters (Vohs, 2007) or other parameters such as intentions attributed to others or education. This heterogeneity is also stressed by Ryan and Deci (2000) who states that individuals 'vary not only in level of motivation (*i.e.*, how much motivation) but also in orientation of that motivation (i.e., what type of motivation). Orientation of motivation concerns the underlying attitudes and goals that give rise to action – that is, it concerns the why of actions'.

Even if they adopt a different viewpoint on intrinsic motivation, Bateman and Crant (2003) acknowledge individual differences in motivational orientation.

Plainly, taking into account the heterogeneity regarding intrinsic motivations is very important for policymaking, insofar as it provides a better description of human behaviour. Unfortunately most papers to date neglect this aspect and, in addition, do not propose alternatives to address the crowding out effect; they merely propose to eliminate monetary incentives.

Our paper addresses these two issues in a theoretical model that (i) allows to take into account the heterogeneity of individuals when faced with the introduction of monetary incentives (ii) suggests a mechanism to reduce the likelihood of getting a crowding-out result by tapping into the heterogeneity of individuals. We assume that monetary incentives matter and change behaviours in predictable directions according to the matching between to whom they are directed *(i.e.* paying the individual versus paying the cause) and the preexisting level of intrinsic motivation of the individual (*i.e.*, low versus high level). Our model can explain a large variety of outcomes stressed in recent empirical studies and has policy relevance by suggesting a new instrument which eventually turns crowding-out into crowding-in (e.g., Beretti et al., 2013).

2 Crowding-out with heterogenous agents: a simple model

This section constructs a simple behavioral model in which it is possible to explore the logical implications of various external monetary incentives on contributions when agents have heterogeneous intrinsic motivations. The model is framed in the environmental realm to fix ideas, but it is nevertheless clear that it could could be applied to many domains where there is a mix of intrinsically motivated individuals and extrinsically motivated ones, such as volunteering or giving[1].

In the kind of situation we analyze in this paper, people know that they affect each other by their decisions, but their interactions are largely anonymous. They clearly don't know the set of strategies of the other individuals, nor do they know their utility functions. Actually they even ignore how many "others" there are. Therefore we prefer to analyze the issue using a "decision-theoretic approach", in connection with the work of Bolle & Otto (2010), rather than with a "game-theoretic approach" as in

[1]For instance, let us consider that intrinsically motivated donators are those who enjoy donating for its own sake. If we study experimental results of dictator games, we get a whole range of individuals from those giving nothing to those giving their own endowments.

Bénabou & Tirole (2006).

There is a continuum of agents of unit mass. Each agent i is endowed with an exogenous income y_i. The decision x_i of agent i is to contribute ($x_i = 1$) or not ($x_i = 0$) to some environmental cause and the opportunity cost of contributing, in monetary terms, is $c(x_i)$. The standard assumption is that this cost is an increasing function: $c(0) < c(1)$. Units are chosen in such a way that $c(0) = 0,\ c(1) = c > 0$. The remainder of the agent's income is affected to some alternative use $c_i = y_i - c(x_i)$.

The conceptual challenge of the present article is most entirely contained in the formulation of the objective function that the agents presumably maximize. Recall that we wish to capture heterogenous intrinsic motivations. And we want to give a role not only to the level of the motivation but also to its orientation. This last aspect in particular means that *procedural* or *frame* consideration is an argument in the agents' objective functions, *i.e.* the same decision x_i performed under two different frames can result in different perceived consequences by the agents. For evidence that people value not only outcomes, but also the procedures or frames that lead to the outcomes see for instance Frey, Benz & Stutzer (2004). Recognition of procedural concerns in agents' choice has recently led to reconsider both the field of *decision theory* (see Salant & Rubinstein, 2008, for an axiomatic analysis of individual choices with frames) and that of *social choice* and *social welfare* (see Suzumura, 1999, Suzumura & Xu, 2001, 2003, Bernheim & Rangel, 2007, 2009, Fleurbaey & Schokkaert, 2013). And, inevitably, one must also reconsider the design of policy instruments, because instruments not only have an effect on x_i, but also because they are part of frames, and frames affect choices directly by themselves. This paper can be seen as a step in that direction.

Let f refers to the frame that agent i faces in a particular choice situation. And assume that agent i is endowed with *decision-relevant* preferences, defined over bundles (x_i, f), numerically represented a *decision utility function* $U^i(x_i, f)$ .We use here the popular distinction between *decision utility* which prompts actions, and *experienced utility,* or hedonic satisfaction[2], which results from actions (see for instance Kahneman, Wakker & Sarin, 1997). It is fairly possible that the same decision x_i made under two different frames f and f', produces different decision utilities: $U^i(x_i, f) \neq U^i(x_i, f')$. Put differently, to each frame f corresponds a particular decision utility function $U^i_f(x_i) \equiv U^i(x_i, f)$.

We propose to study the effects of different frames, with a focus on their interactions with intrinsic

[2]Experienced utility functions would rather incorporate altruism, and capture the public good collectively created by inserting the others' aggregated contribution as an argument in utility functions of the kind $U^i(x_i, x_{-i}, f)$.

motivations. Based on previous studies, Bowles and Polania-Reyes (2012 and references therein) have identified four ways by which those interactions could take place. First, extrinsic incentives may provide information about the person or principal who has chosen the incentives. Second, extrinsic incentives reframe an interaction from one in which effort is required based on moral reasoning to one in which effort becomes a choice because the incentives highlight a possible tradeoff that was not considered previously. Third, extrinsic incentives compromise a control averse individual's sense of autonomy. Fourth, providing extrinsic incentives affect the process by which people learn new preferences.

We consider four distinct frames where all those ways are at play to different degrees: (N) a neutral treatment without external monetary incentives, (A) a treatment where individuals are paid for their contribution x_i, (B) a treatment where agents' decision to contribute is accompanied by a payment directed to a cause supporting the environment (say an relevant association or NGO), (C) a treatment where the agent is offered the choice regarding the orientation of the payment (for himself or for an association). The corresponding decision utility functions are denoted respectively $U_N^i(x_i), U_A^i(x_i), U_B^i(x_i)$, and $U_C^i(x_i)$. Below we make use of behavioral models, using specific funtional forms for each $U_f^i(x_i)$, $f = N, A, B, C$. Beyond those functional forms, the interested reader is invited to check that generalisations of the results of the present paper are possible. The advantage of the simple forms we use is to offer a quick and clear way to highlight the logic we are studying.

2.1 Neutral Treatment (N): pristine altruism alone

In the neutral treatment there is no incentive scheme and altruism is the only intrinsic motivation at work. Assume the decision utility function reads as:

$$U_N^i(x_i) = y_i - c(x_i) + a_i t^N x_i. \tag{1}$$

In expression (1):

- t^N is the marginal "monetarized" benefits produced by the agent's contribution x_i on the other agents. This information parameter is frame-dependent and it is associated here to the situation without external incentives;
- and $a_i \in [0, 1]$ is a parameter that captures an attitude towards the other individuals *via* the en-

vironment, a sort of ecologically-mediated, or 'green', altruistic concern[3]. Those parameters are uniformly distributed on $[0,1]$ and each agent can be identified with a particular point in this interval. At one extreme, Agent 0 with $a_0 = 0$ does not feature any environmental concern; at the other polar case, Agent 1 with $a_1 = 1$ has a strong ecological conscience.

Assume that, for the most altruistic agent, with $a_1 = 1$, the marginal benefit of contributing covers its marginal cost, that is $c(1) - c(0) = c < t^N$. Individuals choose to contribute or not with a view to maximize (1). Hence, agents who settle for zero contributions are those such that:

$$y_i - c(0) > y_i + a_i t^N - c(1) \ ,$$

and the others contribute. Put differently, those in the interval $\left[0, a^N\right[$ where

$$a^N = \left[c(1) - c(0)\right] / t^N = c/t^N, \tag{2}$$

are non-contributors, with a mass a^N, and those in the interval

$$C^N \equiv \left[a^N, 1\right]$$

are contributors. The total number of contributors is

$$1 - a^N \ . \tag{3}$$

2.2 Direct Treatment (A): distorted altruism and moral repugnance

When individuals are paid for their contribution to the environment their decision utility function becomes:

$$U_A^i(x_i) = y_i - c(x_i) + w x_i + a_i t^A x_i - m\left(a_i, w, x_i\right) \ , \tag{4}$$

where w is the monetary payment for participation. The introduction of the payment has two effects as far as intrinsic motivations are concerned.

First, the altruistic motivation is distorted. The idea is that the presence of a monetary transfer acts

[3]The parameter a_i could also represent the degree of altruism or reciprocity.

as a signal of the value of participation (Bolle and Otto, 2010), upon which the agents' marginal benefits of the agents' altruism becomes $a_i t^A$ instead of $a_i t^N$, with $t^A < t^N$. Moreover we assume that this price signal for altruism is at least equal to the payment offered ($t^A \geq w$).

Second, agents who have a concern for the environment now suffer from a *moral repugnance* associated with the fact of being paid for contributing. This aspect is captured by function $m(.,.,.)$ in expression (4). Putting a price onto a territory previously immune to the market forces is one of the list of events that generally spark the yuck factor argument (see for instance Sandel, 2012, and also the discussion about obnoxious markets in Kanbur, 2001). The monetarized value of this psychological "cost" is $m(a_i, w, x_i) \geq 0$. It is natural to assume that the larger the green altruism a_i, or the larger the payment w, and the larger the moral repugnance. To put it formally, $m(a_i, w, x_i)$ is non decreasing in the two first arguments:

$$\begin{aligned} \frac{\partial}{\partial a_i} m(a_i, w, x_i) &\equiv m_a(a_i, w, x_i) \geq 0, \\ \frac{\partial}{\partial w} m(a_i, w, x_i) &\equiv m_w(a_i, w, x_i) \geq 0. \end{aligned}$$

We also assume that the marginal moral repugnance increases when altruism gets larger:

$$\frac{\partial^2}{(\partial a_i)^2} m(a_i, w, x_i) \equiv m_{aa}(a_i, w, x_i) \geq 0 .$$

On the other hand, the mere fact to contribute could mitigate moral repugnance, so $m(a_i, w, x_i)$ is non increasing in the last argument:

$$m(a_i, w, 0) \geq m(a_i, w, 1) .$$

Finally, without altruism, or without payment (when $w = 0$ or/and $x_i = 0$), there is no moral repugnance, therefore

$$m(0, w, x_i) = m(a_i, 0, x_i) = m(a_i, w, 0) = 0 .$$

Agents decide to contribute or not by comparing the levels of utility attached to each possibility. Define

the utility change of contributing:

$$\begin{aligned}\triangle U_A^i(a_i, w) &\equiv U_A^i(1) - U_A^i(0)\,, \\ &= w + a_i\, t^A + c(0) - c(1) + m(a_i, w, 0) - m(a_i, w, 1)\,, \\ &= w + a_i\, t^A - c - \triangle m(a_i, w)\,, \end{aligned} \tag{5}$$

where $\triangle m(a_i, w) \equiv m(a_i, w, 1) - m(a_i, w, 0) = m(a_i, w, 1)$. Function $\triangle U_A^i(a_i, w)$ is supposed to be of class C^1. (meaning that $m(.,.,1)$ as a function of a_i and w is itself C^1.) Notice that increasing the degree of green altruism can have two opposite effects on the utility change. The first effect is positive; it goes through the marginal benefit on others that is more valued by a more altruistic agent. The second effect is negative, because more altruism goes along with a more stringent moral repugnance under direct treatment A.

Contributors are those agents with $\triangle U_A^i(a_i, w) \geq 0$ and non contributors are agents i with $\triangle U_A^i(a_i) < 0$.

Assumption 1. *Assume that:*

$$\lim_{a_i \to 0} \triangle U_A^i(a_i, w) = w - c < 0\,, \tag{6}$$

$$\lim_{a_i \to 1} \triangle U_A^i(a_i, w) = w + t^A - c - \triangle m(1, w) < 0\,, \tag{7}$$

$$\lim_{a_i \to 0} \frac{d}{da_i} \triangle U_A^i(a_i, w) = t^A - m_a(0, w, 1) > 0\,, \tag{8}$$

$$\lim_{a_i \to 1} \frac{d}{da_i} \triangle U_A^i(a_i, w) = t^A - m_a(1, w, 1) < 0. \tag{9}$$

Item (6) of Assumption 1 focuses the analysis to payments w that are not high enough to encourage participation of the less altruistic agents (the payment alone is not enough to compensate the opportunity cost of contributing). This is the most interesting case, because if extrinsic incentives are too strong, no crowding-out effect can occur. Item (7) of Assumption 1 means that for the most altruistic agents, contributing is not optimal because their feeling of altruism towards others, though important, is over-

whelmed by their moral repugnance of being paid. From items (8) and (9) of Assumption 1, given that:

$$\frac{d^2}{(da_i)^2}\triangle U_A^i\left(a_i,w\right)=-m_{aa}\left(a_i,w,1\right)\leq 0,$$

and by the intermediate value theorem, $\exists a_i^*$ such that $t^A-\frac{\partial}{\partial a_i}m\left(a_i^*,w,1\right)=0$. Therefore function $\triangle U_A^i\left(a_i\right)$ has an inverted U shape: it is first increasing, until a_i^*, then decreasing. Assume that $\triangle U_A^i\left(a_i^*\right)>0$. Then there exists a neighborhood

$$C^A\left(a_i^*\right)\equiv[a_i^*-\underline{\varepsilon},a_i^*+\overline{\varepsilon}]\sqsubset[0,1]$$

of contributing agents around a_i^*, *i.e.* $\triangle U_A^i\left(a_i,w\right)\geq 0,\forall a_i\in C^A\left(a_i^*\right)$. Note that $C^A\left(a_i^*\right)$ is a *proper subset* of $[0,1]$, for all the elements of $[0,1]$ do not belong to $C^A\left(a_i^*\right)$; in particular, because of parts (6) and (7) of Assumption 1, agents a_0 and a_1 are not contributors. It is interesting to emphasize the peculiarity of this treatment. The choice to contribute can be explained by two intrinsic motivations of different natures: a degree of altruism sufficiently high or a moral repugnance not too strong (precisely in the most altruistic agents).

For future reference, let us denote:

$$\begin{aligned}\underline{a}^D &= a_i^*-\underline{\varepsilon}\,,\\ \overline{a}^D &= a_i^*+\overline{\varepsilon}\,.\end{aligned}$$

the agents who, among those who contribute, have the lowest and largest altruism respectively. By definition, those two values solve the equation:

$$\triangle U_A^i\left(a_i,w\right)=\ w+a_i\,t^A-c-\triangle m\left(a_i,w\right)=0. \tag{10}$$

Does crowding-out necessarily occur? To answer this question, one has to compare the mass of $1-a^N$ of contributors under the neutral treatment with the mass $\overline{a}^D-\underline{a}^D$ of contributors under treatment A. Unless more structure is given to the moral repugnance function $\triangle m\left(a_i,w\right)$, equation (10) cannot be solved explicitly for $\underline{a}^D$ and $\overline{a}^D$. Still, important qualitative pieces of information can be obtained. The possibility of crowding-out depends on the relative position of a^N with respect to $\underline{a}^D$ and $\overline{a}^D$. The answer

is ambiguous when $a^N \in \left[\underline{a}^D, \overline{a}^D\right]$ and when $a^N > \overline{a}^D$, but there is crowding-out for sure when $a^N \leq \underline{a}^D$. All those situations can be associated with particular conditions on parameters. Since $\triangle U_A^i\left(a_i\right)$ has an inverted U shape and takes on positive values around a_i^*, by construction $a^N \in \left[\underline{a}^D, \overline{a}^D\right]$ if and only if:

$$\triangle U_A^i\left(a^N, w\right) \geq 0.$$

Using (5) and the fact that $a^N = c/t^N$ (see (2)):

$$\triangle U_A^i\left(a^N, w\right) \geq 0 \quad \Leftrightarrow \quad w + \frac{c}{t^N} * t^A - c - m\left(\frac{c}{t^N}, w, 1\right) \geq 0\,. \tag{11}$$

By the same logic, if $a^N < \underline{a}^D$ or if $a^N > \overline{a}^D$, then necessarily:

$$\triangle U_A^i\left(a^N, w\right) < 0 \quad \Leftrightarrow \quad w + \frac{c}{t^N} * t^A - c - m\left(\frac{c}{t^N}, w, 1\right) < 0\,. \tag{12}$$

The last inequality means that, for agent a^N, the payment alone does not provide sufficient incentives to compensate the cost of moral repugnance and the decrease in altruistic motivation following the change of benefits on others from t^N to t^A. This is consistent both with a too low value for w and with a too high value (recall that moral repugnance increases with w). This assessment of the weakness of monetary incentives is not absolute, but relative to parameters t^N and t^A. So, rewriting equality (12):

Definition 1 ((N/A)-weak extrinsic incentive)**.** *Extrinsic incentives are (N/A)-weak when:*

$$w < c\left(\frac{t^N - t^A}{t^N}\right) + m\left(\frac{c}{t^N}, w, 1\right).$$

The above reasoning has therefore established:

Proposition 1. *If $a^N < \underline{a}^D$, incentives are (N/A)-weak and there is crowding out.*

A last question is about the interactions between internal and external incentives. It is generally considered that the phenomenon of crowding-out gets weaker as external (monetary) incentives gets stronger. Let us analyze the effect of increasing the external incentive w on the upper and lower bounds of $C^A\left(a_i^*\right)$, *i.e.* on $\underline{a}^D$ and $\overline{a}^D$. By equation (10) and the implicit function theorem, using the properties that $t^A - m_a > 0$ until a_i^* and $t^A - m_a < 0$ after a_i^*, and under the assumption that moral repugnance

increases less than proportionally with the external incentive, $m_w < 1$, we can conclude:

$$\begin{aligned} \left.\frac{da}{dw}\right|_{a=\underline{a}^D} &= -\frac{1-m_w\left(\underline{a}^D,w,1\right)}{t^A-m_a\left(\underline{a}^D,w,1\right)}<0\,, \\ \left.\frac{da}{dw}\right|_{a=\overline{a}^D} &= -\frac{1-m_w\left(\overline{a}^D,w,1\right)}{t^A-m_a\left(\overline{a}^D,w,1\right)}>0\,. \end{aligned}$$

Therefore, as w increases the mass of contributors $C^A\left(a_i^*\right)=\left[\underline{a}^D,\overline{a}^D\right]$ gets wider. But it is important to notice that a crowding-out is not a necessary consequence of the direct treatment. A necessary condition for *crowding in* is $\underline{a}^D < a^N$. This may well happen if the monetary payment w is high enough. But this condition is not sufficient, because there is a mass of highly altruistic agents, $1-\overline{a}^D$, who do not participate. By continuity, when w increases so that $\underline{a}^D$ decreases and falls below a^N, and if:

$$\left|\left.\frac{da}{dw}\right|_{a=\underline{a}^D}\right| < \left|\left.\frac{da}{dw}\right|_{a=\overline{a}^D}\right|\,,$$

that is if at the margin the increase of lower-end contributors is less than the decrease of upper-end non contributors, there is a continuum of values for w consistent with crowding-in, even though the less altruistic agent who participates under the direct treatment is less altruistic than the less altruistic agent who participates in the neutral treatment.

To summarize,

Proposition 2. *Under Assumptions (8), (9), and when moral repugnance increases less than proportionally with the external incentive, $m_w < 1$, the stronger the external incentive w, the weaker the crowding-out effect, if any, under the direct treatment.*

2.3 Indirect Treatment (B): distorted altruism alone

Under this design the payment is no longer given to contributors; rather it is directed to a cause supporting the environment, for example a related association or NGO. Hence individual i no longer bears the cost of moral repugnance, but of course his altruistic motivation is activated, for participation still generates a benefit to the environment.

The decision utility functions are now:

$$U_B^i(x_i) = y_i + a_i t^B x_i - c(x_i) \ ,$$

where t^B is the corrected marginal percieved benefits on the environment.

Regarding the fact that a payment is directed to the cause supported by the individual, we can reasonably consider that the perceived benefit on the environment of his participation is higher than when the same amount of money is directed to the individual's pocket, because the chosen destination, by its very nature, reinforces the belief of the agent on the presence of high environmental values, or because the association is more efficient than individuals in transforming a given amount of contributions in environmental gains. An assumption on parameters consistent with that view is $t^A \leq t^B$.

Hence, agents who settle for zero contributions are those such that:

$$y_i - c(0) > y_i + a_i \, t^B - c(1),$$

and the others contribute. Put differently, there is an interval $\left[0, a^I\right[$, where

$$a^I = \frac{c}{t^B}, \tag{13}$$

of non-contributors, and an interval

$$C^B \equiv \left[a^I, 1\right]$$

of contributors. The total number of contributors is

$$1 - a^I \ . \tag{14}$$

Compared to the neutral treatment there is crowding-out when $t^B \leq t^N$, because the mass of contributors has shrunk (compare expression (14) with expression (3)). This non ambiguous result is rather intuitive: there is no direct own benefit and the estimation of the benefits on others has been cut down ($t^B \leq t^N$), so the incentives to participate are weaker compared to the neutral treatment.

However, the comparison with the direct treatment is more subtle. It is worth noting that there is a

mass of agents near a_1 who contribute under the indirect treatment and who do not contribute under the direct treatment. In two cases, when $a^I \in \left[\underline{a}^D, \overline{a}^D\right]$ and when $a^I > \overline{a}^D$, we cannot state which policy better encourages participation. But the indirect treatment out-performs the direct treatment for sure when $a^I < \underline{a}^D$. Again all those situations can be associated with particular conditions on parameters. For $\underline{a}^D \leq a^I \leq \overline{a}^D$, a necessary and sufficient condition on the fundamentals of the model derives from the observation that $\triangle U_B^i\left(a^I, w\right) = \triangle U_B^i\left(c/t^B, w\right) \geq 0$ in such a situation. Or equivalently, using (5):

$$w \geq c * \frac{t^B - t^A}{t^B} + m\left(\frac{c}{t^B}, w, 1\right) . \tag{15}$$

When this condition is not met, a necessary condition is obtained for $a^I < \underline{a}^D$:

Definition 2 ((B/A)-weak extrinsic incentives)**.** *Extrinsic incentives are (B/A)-weak when:*

$$w < c * \frac{t^B - t^A}{t^B} + m\left(\frac{c}{t^B}, w, 1\right) .$$

When $a^I < \underline{a}^D$, incentives are (B/A)-weak and crowding-out is unambiguously less important under the indirect treatment. Intuitively, even if there are no monetary rewards for the agents under policy B, altruistic motives are less corroded than under the direct treatment and, in addition, the extrinsic motivation is not strong enough under policy A to compensate the weaker altruistic motivation and moral repugnance.

In a nutshell:

Proposition 3. *When the estimations of the benefits on others are such that $t^A \leq t^B \leq t^N$, if $a^I < \underline{a}^D$ the extrinsic incentives are (B/A)-weak and the indirect treatment unambiguously mitigates the crowding-out effect compared to the direct treatment. When extrinsic incentives are not (B/A)-weak, or when $a^I > \overline{a}^D$, the ability of the indirect treatment to improve participation compared to the direct treatment is ambiguous. However, in any case, participation under the indirect treatment is never larger than under the neutral treatment, $a^N \leq a^I$.*

2.4 Choice Treatment (C): auto-selection of motivations

Under this treatment, individuals can choose whether the payment is directed to themselves or to an environmental association. Giving the choice to individuals (keeping the reward for themselves or giving it to the 'environmental cause') could motivate a wider set of individuals, possibly leading to a higher overall contribution[4].

In a sense, by choosing the target of the payment individual i chooses which decision utility function to activate. Then, agent i's decision utility function U_C^i exists in two expressions:

- it is:

$$U_C^i(x_i) = U_A^i(x_i) = y_i - c(x_i) + wx_i + a_i t^A x_i - m\left(a_i, w, x_i\right) \ ,$$

  when the payment is direct.

- and it is:

$$U_C^i(x_i) = U_B^i(x_i) = y_i + a_i \, t^B x_i - c(x_i) \ ,$$

  when the payment is indirect.

Does the choice treatment minimize the countervailing effect of external incentives?

Notice first that the utility attached to non participation is the same, whatever the chosen target:

$$U_C^i(x_i) = U_N^i(0) = U_A^i(0) = U_B^i(0) = y_i.$$

But the utility derived from participation differs according to the target of the payment. Agents who increase their utility by contributing are those who belong to at least one of the sets of contributors previously identified. Clearly, the set C^C of contributors under the choice treatment is the union of the two sets of contributors of each separate treatment, *i.e.* $C^C = C^B \cup C^A$. The set C^C encompasses agents a_i such that $\triangle U_B^i\left(a_i\right) \geq 0$, and/or $\triangle U_B^i\left(a_i\right) \geq 0$ and, therefore, the choice treatment promotes participation as least as much as the two policies A & B separately do. But more precision can be added.

We will keep on assuming that estimations of the benefits of altruism are such that $t^A \leq t^B \leq t^N$. Even under this assumption, several configurations for the different sets of contributors are possible:

[4]Moreover, in addition to obvious intuitive reasons based on empirical evidence, we argue that people enjoy the possibility of choosing by themselves, even at a cost (Benz et al., 2004; Frey and Stutzer, 2005).

Case 1 A first case is when $a^N \leq a^I < \underline{a}^D$, so the extrinsic motive is both (N/A)-weak and (B/A)-weak (definitions 1 and 2). Then the different sets of contributors are such that:

$$C^A \sqsubset C^B = C^C \sqsubseteq C^N.$$

Contributors under the choice treatment are exactly those who contribute under the indirect treatment and they are not more numerous than those who contribute under the neutral treatment.

Case 2 A more interesting case is when the monetary payment is sufficiently important to produce the following ranking $a^N \leq \underline{a}^D < a^I \leq \overline{a}^D$, that is the extrinsic motive is (N/A)-weak but it is not (B/A)-weak. The sets of contributors are then in the following configuration:

$$C^A, C^B \sqsubset C^C \sqsubseteq C^N.$$

Contributors under the choice treatment are more numerous than those who contribute under any of the two separate treatments. But the choice treatment does not perform any better than the neutral treatment.

Case 3 The most interesting case is when the monetary incentives are pushed slightly further so that $\underline{a}^D < a^N < a^I \leq \overline{a}^D$. The sets of contributors are such that:

$$C^A, C^B \sqsubset C^N \sqsubset C^C.$$

This is a case featuring crowding-out under each separate treatment, but there is crowding-in under the choice treatment. This possibility occurs because several intrinsic motivations exists and because agents are heterogenous. As a results those who contribute are not necessarily identical across treatments and, even more, C^A is neither a proper subset of C^B nor a proper subset of C^N. The corresponding necessary and sufficient conditions on parameters have been identified in (11) and (15). They must be imposed simultaneously, as a new assumption:

Assumption 2 (Conditions for crowding-in)**.**

$$w \geq c * \frac{t^N - t^A}{t^N} + m\left(\frac{c}{t^N}, w, 1\right) \quad \textit{and} \quad w \geq c * \frac{t^B - t^A}{t^B} + m\left(\frac{c}{t^B}, w, 1\right).$$

Case 4 Finally when $\underline{a}^D < a^N < \overline{a}^D \leq a^I$. It is not possible to conclude - without further information on the moral repugnance function - about the extent of the crowding-out phenomenon, if any, because there is a mass of agents characterized by intermediate degrees of altruism in the interval $\left]\overline{a}^D, a^I\right[$ who are not contributors. However, this situation is discarded when the extrinsic motivation is not t^B-weak.

To summarize, the choice treatment combines the incentive effects of both the direct and indirect treatments:

Proposition 4. *Let Assumption 1 holds and assume also $t^A < t^B \leq t^N$. Participation under policy C (choice treatment) is at least as large as under policies A & B. The choice treatment even results in crowding-in, although there is crowding-out under policies A & B, if and only if Assumption 2 is satisfied.*

3 Conclusion

We made a strong case for how different intrinsic motivations among agents can play an instrumental role in explaining the effectiveness of introducing monetary incentives. We have formalized how the heterogeneity of intrinsic motivations among agents impacts their reactions to the introduction of monetary incentives. We showed that overall results supporting (or not) an undesired crowding-out effect can occult a more complex reality where some individuals contribute thanks to these additional monetary incentives while others reduce their contributions. Moreover, we proposed a new instrument which taps into agents' heterogeneity in order to suppress, or at least to reduce, the risk of crowding-out effect. This instrument avoids a 'one-size-fits-all' policy and allows agents to self-select the most relevant arrangement. A considerable advantage of our mechanism is that it does not require that policy makers have an extensive knowledge about the various levels of intrinsic motivations of agents.

Theoretically, the proposed instrument respects the freedom of choice of individuals. Indeed, they decide about the final use of the received monetary incentives. Nevertheless, we are conscious that the possibility of choice might also strongly vary with the framing of the task and could change the social behaviour leading to different normative expectations. In line with the traditional maxim stressing that

the evil is in the details, we encourage a careful framing of real-world instruments by pre-testing their various versions in pilot experiments.

References and Notes

[1] Bénabou R., Tirole J., 2006, *Incentives and Prosocial Behavior*, American Economic Review. 96, 1652-1678.

[2] Benz M., Frey B.S., Stutzer A., 2004, *Introducing Procedural Utility: Not Only What, but Also How Matters*, Journal of Institutional and Theoretical Economics JITE. 160, 377-401.

[3] Beretti A., Figuières C., Grolleau G., 2013, *Using Money to Motivate Both Saints and Sinners: a Field Experiment on Motivational Crowding-Out*, Kyklos, 66/1, 63-77.

[4] Bernheim D., Rangel A., 2007. *Toward Choice-Theoretic Foundations for Behavioral Welfare Economics*. American Economic Review, 97(2): 464-470.

[5] Bernheim D., Rangel A., 2009, *Beyond revealed preference: choice theoretic foundations for behavioral welfare economics*. Quarterly Journal of Economics, 124(1), 51-104.

[6] Bolle F., Otto P.E., 2010, *A Price Is a Signal: on Intrinsic Motivation, Crowding-out, and Crowding-in*, Kyklos. 63, 9-22.

[7] Bowles S., 2008, *Policies Designed for Self-Interested Citizens May Undermine "The Moral Sentiments": Evidence from Economic Experiments*, Science. 320, 1605.

[8] Fleurbaey M., Schokkaert E., 2013, *Behavioral welfare economics and redistribution*, American Economic Journal: Microeconomics 5(3): 180-205.

[9] Frey B.S. , Oberholzer-Gee F., 1997, *The Cost of Price Incentives: An Empirical Analysis of Motivation Crowding- Out*, The American Economic Review. 87, 746-755.

[10] Frey B.S., Benz M., Stutzer A., 2004, *Introducing procedural utility: not only what, but also how matters*, Journal of Institutional and Theoretical Economics, 160, 377–401.

[11] Frey B.S., Stutzer A., 2005, *Beyond Outcomes: Measuring Procedural Utility*, Oxford Economic Papers. 57, 90-111.

[12] Goette L., Stutzer A. & Frey B.M., 2010, *Prosocial Motivation and Blood Donations: A Survey of the Empirical Literature*. Transfusion Medicine and Hemotherapy, 37(3), 481-502.

[13] Kanbur R., "On Obnoxious Markets", 2001. Revised version published in Stephen Cullenberg and Prasanta Pattanaik (editors), Globalization, Culture and the Limits of the Market: Essays in Economics and Philosophy. Oxford University Press, 2004.

[14] Kelly D., 2011, *Yuck! The Nature and Moral Significance of Disgust*, MIT Press.

[15] Gneezy U., Meier S., Rey-Biel P., 2011. *When and Why Incentives (Don't) Work to Modify Behavior*, Journal of Economic Perspectives, 25(4), 191-210.

[16] Kahneman D., Wakker P., Sarin R., 1997. *Back to Bentham? Explorations of Experienced Utility*, The Quarterly Journal of Economics, 112(2), 375-405.

[17] Mellström C., Johannesson M., 2008, *crowding-out in Blood Donation: Was Titmuss Right?*, Journal of the European Economic Association. 6, 845-863.

[18] Reeson A.F., Tisdell J.G., 2008, *Institutions, motivations and public goods: An experimental test of motivational crowding*, Journal Of Economic Behavior and Organization. 68, 273-281.

[19] Salant Y., Rubinstein A., 2008, *(A, f): Choice with Frames*, Review of Economic Studies (2008) 75 (4): 1287-1296

[20] Sandel M., 2012, What money can't buy: the moral limits of markets, Allen Lane, A/L PENG PRESS.

[21] Suzumura K., 1999. Consequences, opportunities, and procedures, *Social Choice and Welfare*, 16(1), 17-40.

[22] K. Suzumura, Xu Y., 2001, *Characterizations of consequentialism and non consequentialism*, Journal of Economic Theory, 101, 423–436.

[23] Suzumura K., Xu Y., 2003, *Consequences, opportunities, and generalized consequentialism and non-consequentialism*, Journal of Economic Theory, 111(2), 293-304.

[24] Titmuss R.M., 1970, *The gift relationship: From human blood to social policy*, Allen and Unwin.

[25] Tor Helge Holmås, Egil Kjerstad and Hilde Lurås, Odd Rune Straume, 2010, Does monetary punishment crowd out pro-social motivation? A natural experiment on hospital length of stay", Journal of Economic Behavior & Organization, 75, 261-267.