Center
for
Economic Research

No. 2000-71

THE LIMIT OF PUBLIC POLICY: ENDOGENOUS PREFERENCES

By Oren Bar-Gill and Chaim Fershtman

August 2000

ISSN 0924-7815

The Limit of Public Policy: Endogenous Preferences

Oren Bar-Gill[*] and Chaim Fershtman[**]

August 2000

Abstract: In designing public policy it is not enough to consider the possible reaction of individuals to the chosen policy. Public policy may also affect the formation of preferences and norms in a society. The endogenous evolution of preferences, in addition to introducing a conceptual difficulty in evaluating policies, may also eventually affect actual behavior. In order to demonstrate the implications of endogenous preferences on the design of optimal public policy, we present a model in which a subsidy policy is set to encourage contributions towards a public good. However this policy triggers an endogenous preference change that results in a lower level of contribution towards the public good despite the explicit monetary incentives to raise that level.

[*] The Eitan Berglas School of Economics, Tel Aviv University, and Harvard Law School.
[**] The Eitan Berglas School of Economics, Tel Aviv University, Tel Aviv 69978, Israel and CentER, Tilburg University.

"Social scientists – especially economists – evaluate policies and institutions by examining behavioral responses to incentives, with values, habits, and social norms taken as given and beyond analysis and the reach of public policy" Aaron (1994)

1. Introduction

The axiom of exogenous preferences is one of the cornerstones of economic modeling. Models typically begin with the specification of players' preferences and then proceed to analyze market interactions. The exogenous preferences assumption reflects the general approach of "taking individuals as they are," although it has, at the same time, considerably simplified the task of economic theory.[1] Models that examine public policy are no different. The focus of these models is on the effect of public policy on the outcomes of the market interaction assuming that preferences are exogenously given. The possible effect of public policy on the formation of preferences, values and norms of behavior is largely ignored.[2]

In recent years, we have seen a growing literature that studies the endogenous formation of preferences. The basic approach in this literature is that preferences evolve over time, depending on the players' actions and the outcomes of market interactions. The preference dynamics may be the outcome of imitation, education, or other forms of cultural transmission dynamics. Preference dynamics have been studied so far mainly using an evolutionary approach.[3] In this approach, a fitness criterion

---

1 It is interesting to quote Becker's view on this issue: "… all human behavior can be viewed as involving participants who maximize their utility from a stable set of preferences..." (1976).

2 For more on this problematic approach see Bowles (1998), who wrote: "Markets and other economic institutions do more than allocate goods and services: they also influence the evolution of values, taste, and personalities."

3 The evolutionary approach does not necessarily advocate a biologically based preference dynamics although recent studies have shown the existence of genes that determine preference attributes like risk aversion. While the setup is based on evolutionary dynamics, the mechanism that governs these dynamics may be rooted in behavior such as imitation and education. For evolutionary models that endogenize preferences see Basu (1995), Bester and Güth (1998), Dekel and Scotchmer (1999), Fershtman and Weiss (1997,1998), Huck and Oechssler (1999), Koçkesen et al. (2000), Robson (1996), and Rogers (1994).

replaces preferences as the exogenously given primitive of the model. Within a population, the frequency of individuals with certain preferences increases if their fitness is above the average fitness in the population.[4] Another possible approach to modeling preference dynamics is cultural transmission (e.g., Cavalli-Sforza and Feldman (1981) and Bisin and Verdier (1998)), in which the effects of imitation and education on preference dynamics are explicitly modeled. In both approaches, however, the preference dynamics imply that the preference profile at period *t+1* depends on the preference profile at period *t*, on the outcome of market interactions at period *t,* and possibly on the individuals' efforts to transmit their preferences to the next generation.

Preference dynamics introduces a direct link between public policy and the formation of preferences. Public policy changes the outcomes of market interactions and thus affects the evolution of future preferences profiles.[5] Thus, in designing an optimal public policy, it is not enough to consider the possible reaction of individuals to that policy; we should take into account as well the effects of different policies on the formation of preferences and subsequently on behavior.

The relationship between public policy and preferences introduces conceptual difficulties in studying (or even defining) optimal public policies. The standard economic modeling approach is to define the optimal policy with respect to an optimality criterion (such as maximum total welfare) which ranks market outcomes given an exogenously specified profile of preferences. An optimal policy is, therefore,

---

[4] Clearly, the discussion regarding the appropriate assumptions about preferences is replaced in this model by the discussion about the appropriate fitness criterion. This is an important discussion which is, however, beyond the scope of this paper. Social factors may enter into the fitness criterion but, for simplicity's sake, we will adopt in our example the assumption that the fitness criterion is a monetary payoff.

[5] In some cases the effect may be only transitory. But one cannot exclude (as we will demonstrate in our model) that steady state preferences also depend on public policy.

one that induces a market outcome that is not dominated (in terms of the optimality criterion) by a market outcome induced by another public policy. However, when public policy affects the evolution of preferences, the above selection procedure is no longer valid. The optimality condition may rank outcomes only for given preferences. Moreover, ranking profiles of preferences or comparing outcomes for different preferences is known to be a problematic or even an impossible task. Even if these conceptual problems can be avoided, endogenous preference formation may induce time consistency problems.[6] The policy that was optimal for one preference profile may induce a new preference profile for which the policy is no longer optimal.

In order to demonstrate the implications of endogenous preferences on the design of public policy, we consider a simple example in which contributions towards a public good are encouraged by direct monetary subsidization. We adopt the endogenous preference setting of Fershtman and Weiss (1997), into which we introduce a public good term and study the effects of subsidization on the preference dynamics.[7] Individuals in this model are pair-wise matched and play a Prisoner's Dilemma-like game. The players' actions affect their direct payoffs in the game and the aggregate action contributes toward a public good they all commonly enjoy. Furthermore, the players may also care about their social status, which is determined by their relative contribution to the public good. However, this concern for social status is not necessarily shared by all players. We do not impose any preference profile; rather, we assume an evolutionary selection process that determines the profile endogenously. We then assume that a subsidy policy is used in order to promote the accumulation of the

---

[6] see also Peleg and Yaari (1973).

[7] The main concern of Fershtman and Weiss (1997) was indeed to show that even when the evolutionary process is governed by a fitness function that takes only monetary payoffs into

public good. A short run analysis (assuming given preferences) indicates that such a subsidy indeed increases the equilibrium level of the public good. In the long run, however, the subsidy policy induces a shift in the distribution of preferences, reducing the number of socially minded individuals. Such a shift reduces the social incentives as well as the proportion of the population that cares about them. Consequently, the subsidy policy results in a *lower* level of the public good as the greater monetary incentives do not offset the disappearance of the social incentives.

The importance of social rewards in providing incentives or compensation for individuals who perform activities with positive externalities was already suggested by Arrow (1971). Is society free to use any level of "honor" and "status" as compensation for social services? This issue was addressed in Fershtman and Weiss (1998), who showed that (i) some social activities cannot be induce by social rewards and (ii) when social rewards are overused, they cease to be effective because preferences may endogenously change, thereby reducing the emphasis individuals place on those social rewards. In this paper, however, we emphasize the possible limits of standard monetary incentives in inducing activities with social benefits. The mechanism that explains this phenomenon is one where the use of monetary incentives triggers an endogenous preferences shift. This shift implies not simply lower social incentives, but also a smaller proportion of individuals who care about those particular social rewards.

The above result resembles a well-known argument in social psychology. In a controversial paper, Titmuss (1970) argued that allowing payments for blood donations will result in a lower level (and even in a lower quality) of donation.[8] This hypothesis suggests that in some circumstances, monetary rewards crowd out intrinsic motivation

account, the evolutionary stable preference profile may be such that some or all individuals have social preferences.

like civic duty. The problem of the "crowding out effect" has been studied since then by cognitive social psychologists[9] as well as by economists.[10] While our model predicts similar outcomes, the underlying phenomenon is quite different. The emphasis in the psychological and experimental economics literature is on the effect of monetary incentives on the perception of people regarding the social rewards themselves. That is, if a price is placed on a blood donation, then donating blood is not necessarily such a noble act nor a civic duty. Our model, on the other hand, focuses on the possibility that monetary incentives may induce a change in the underlying preference profile.[11] The outcome of such a preference shift may indeed change the relative importance of intrinsic motivation and of extrinsic monetary rewards.

2. Sketch of the Argument

Consider a population of size N in which individuals may have different preferences. Let $P \equiv \{p_1, \ldots, p_n\}$ be the set of possible preferences (or types). The distribution of types is given by $q \equiv (q_1, \ldots, q_n) \in Q$, such that q_i is the percentage of individuals in the population with preferences p_i. Preferences are defined over a set of possible outcomes, Φ. We define an outcome broadly so that it may consist of an

---

[8] See also Solow (1971) and Arrow (1972) for a critical review of this argument. Both these papers argued that monetary incentives should be simply added to the intrinsic incentives.

[9] See Deci (1971), Deci (1975) and for a survey of this literature see Lane (1991).

[10] See for example Frey (1994), Frey et al. (1996) and Frey and Oberholzer-Gee (1997) for a detailed examination of the crowding-out effect, and Fehr, Gachter and Kirchsteiget (1996), Fehr and Gachter (1998), and Gneezy and Rustichini (2000a,b) for a recent experimental study on the effectiveness of monetary incentives.

[11] The change in preferences may indirectly influence the value placed on the social rewards. For example, if fewer people care about status, the value of status might decline.

allocation of goods as well as of non-monetary elements that induce some social ranking (for example, social status).

Given their preferences, individuals are engaged in a market game in which they need to choose an action $x \in X$. The government may intervene in this market game by formulating a policy that induces a system of incentives or rewards. Let $G \equiv \{g_1, \ldots, g_M\}$ be the set of possible government policies. The players' actions $(x_1, \ldots, x_N)$, together with the government policy g_j, determine the outcome of the market game.

The optimal action of each individual depends on his preferences p_i, on the distribution of preferences in the population *q*, on the actions taken by other individuals and on the government's public policy g_j.[12] For any given public policy g_j and distribution of preference types *q*, we assume that there is a unique equilibrium of the market game and that the equilibrium action chosen by an individual of type *i* is given by $x_i = x(p_i, q, g_j)$. The equilibrium outcome of the market game is denoted as $f(q, g_j) \in \Phi$.[13]

The basic approach of the theory of optimal public policy included the following steps: (i) it adopts an optimality criterion F[14] that, for every given preference profile *q*, induces a (possibly partial) ranking, F_q, over the set of outcomes, Φ; and (ii) given the distribution of preferences *q*, it selects an optimal policy, $g^* \in G$, such

[12] Note that this optimal action may depend on the preferences of other individuals and not just on their actions. For example, in order to allow for preferences with a social element like status individuals may care about their status only if other individuals in their community view status as an important factor (see the discussion in Fershtman and Weiss (1997) and in section 3).

[13] Clearly, existence and uniqueness of the market equilibrium is not guaranteed but, in order to sketch our argument, we will avoid these issues and assume both existence and uniqueness.

[14] The optimality criterion can be, for example, maximum welfare, overall monetary payoffs, Pareto optimality, and so on.

that there is no other public policy $g \in G$ for which $f(q,g)$ dominates $f(q,g^*)$ in terms of the ranking F_q, that is, $F_q(f(q,g^*)) \geq F_q(f(q,g))$ for all $g \in G$.

The stated public policy selection procedure is the standard framework used in the public policy literature. This procedure is well defined under the assumption that preferences are exogenously given, i.e., q is given and not affected by the policy choice g.

The basic structure of the endogenous preferences approach entails that the preference distribution at period *t+1* depends on the preference distribution at period *t* and on the outcome of the market game at period *t*, i.e. $q_{t+1} = D(q_t, f)$. The function $D(\cdot,\cdot)$ is the population dynamics that may be induced by a process of imitation, learning, evolution or cultural transmission. Since the outcome of the market game is affected by the existing public policy, the link between policy and the evolution of preferences is established; hence, $q_{t+1} = D(q_t, f(q_t, g))$.

For every initial distribution of preferences q_0, and public policy g, we can define $q^f(q_0, g)$ as the limit of the above preference dynamics process.[15] The steady state condition is:

$$(1) \quad q^f(q_0, g) = D(q^f(q_0, g), f(q^f(q_0, g)))$$

A preference profile q is stable under the public policy g if there is a small neighborhood of q such that for every q' in this neighborhood, $q^f(q', g) = q$.

When public policy affects preferences, the task of identifying an optimal policy becomes quite problematic. First, there is the familiar conceptual problem of evaluating the implication of a preference change. There are no absolute criteria that can determine

[15] Clearly, this limit does not necessarily exist, but we ignore this issue for now and assume a convergence of the population dynamics process.

if individuals are better off with one set of preferences or another. We can think about an individual with two different sets of preferences as two different individuals but interpersonal comparison of utility is known to be dubious.

When preferences are exogenous, public policy affects only the market equilibrium outcomes. The social ranking F_q, which is derived from the exogenous preference profile and the optimality criterion F, is not a function of g. When public policy changes the population's preference profile from q_t to q_{t+1}, this effect carries over immediately to the ranking, and F_{q_t} becomes $F_{q_{t+1}}$. This raises the question of whether the outcome of the market interaction should be evaluated by the "old" or the

One may also consider the following time consistency problem. Let q_0 be the initial preference profile. Given F_{q_0}, the optimal policy g^* is determined such that $F_{q_0}\left(f(q_0, g^*)\right) \geq F_{q_0}\left(f(q_0, g)\right)$ for all $g \in G$. But when preferences evolve endogenously, the outcome induced by the policy g^* changes the underlying preferences profile to $q_1 = D\left(q_0, f(q_0, g^*)\right)$. This new preference profile results in a new ranking, F_{q_1}, for which g^* is not necessarily optimal, i.e. there is a policy $g \in G, g \neq g^*$ such that $F_{q_1}\left(f(q_1, g)\right) > F_{q_1}\left(f(q_1, g^*)\right)$. A similar consistency problem may arise when we restrict the evaluation of the effect of the public policy to the limit preferences induced by g^*, i.e., $q^f\left(q_0, g^*\right)$. The policy g^*, which is optimal for the initial preference profile q_0, is not necessarily optimal for the ranking F_{q^f}, which corresponds to the limit preferences $q^f\left(q_0, g^*\right)$.

3. Subsidizing a Public Good

We follow Fershtman and Weiss (1997), hereinafter FW, and consider a setting in which preferences are endogenously determined by an evolutionary process. We introduce into that model a public good term and investigate the effects of subsidization on individual behavior, the formation of preferences, and the overall equilibrium level of the public good.

3.1 The Market Interaction

Consider a society with a large number N of individuals. In every period, individuals are pairwise matched and play a Prisoners' Dilemma-type game. Each player in this game needs to choose an effort level e_i from the set $\{0,1\}$. Let $\Pi_i(e_i, e_j)$ be the direct monetary payoffs of player i, who is matched with player j. The values of $\Pi_i(.,.)$ are given in the following payoff matrix:

		player j: 0	player j: 1
player i	0	α , α	β , δ
	1	δ , β	γ , γ

Fig. 1: The Payoff Matrix

where $b > g > a > d$. We further assume that $a - d > b - g$.

In addition to the direct payoff, the players' overall efforts contribute toward a public good which they all commonly enjoy. Let $\hat{e}$ be the total amount of effort in the

[16] One way to avoid this problem and to use the standard optimality criterion is to adopt a multiple

population and let *E(.)* be a public good term such that *E(.),E'(.) > 0.* Player *i*'s overall payoffs are: [17]

$$(2) \quad m_i(e_i, e_j, \hat{e}) \equiv \Pi_i(e_i, e_j) + E(\hat{e}) .$$

We assume that *N* is sufficiently large such that the effect of e_i on $\hat{e}$ is negligible and each player views $\hat{e}$ as fixed. Like FW, we assume that the chosen effort level determines, besides the above monetary payoffs, also the individuals' social status. Players, however, do not necessarily care about status. Some may simply maximize their economic payoffs (2), while others may value a high social status as well. We do not impose any preference profile but derive it endogenously. For simplicity, we allow for only two types of preferences, players that care about their social status and players who totally disregard it. Denoting the social reward by Σ, the utility of player *i* is:

$$(3) \quad U_i = m_i + p_i \Sigma_i$$

where $p_i \in \{0,1\}$ is the preference parameter. Individuals with $p_i=1$, hereinafter type 1 individuals, care about their social status, whereas individuals with $p_i=0$, hereinafter type 0 individuals, do not care about social status. Let $q \in [0,1]$ denote the proportion of type 1 players in the population.

When effort is positively correlated with status, social rewards encourage individuals to contribute toward the public good (see Arrow (1971) for a discussion on the role of social rewards). Letting the average effort e^a, $e^a = \hat{e}/N$, represent the social norm, we assume that status (positive or negative) is conferred upon an individual according to her performance relative to the social norm. We further assume

selves approach. In such a case, it is possible to evaluate the future market outcomes from the point of view of today's preferences. See for example Peleg and Yaari (1973).

[17] To simplify calculations we assume that the public good term enters additively.

that only socially minded individuals can confer status on others. Consequently, the social reward to individual *i* is given by

$$(4)\quad \Sigma_i \equiv qs(e_i - e^a)$$

where s is an exogenously given marginal social reward parameter.

Substituting (2) and (4) into (3), we obtain the following expression for the utility of player *i* -

$$(5)\quad U_i(e_i, e_j, e^a, q) = \Pi_i(e_i, e_j) + p_i qs(e_i - e^a) + E(\hat{e})$$

We assume that the players' types are fully observable.[18] We can now derive the equilibrium actions in the above game. Since there is a large number of players, individual players do not affect the public good term $E(\hat{e})$ which, can therefore be ignored in considering the game between each pair of players.[19]

When the two players are of type 0, the game is represented by the payoff matrix in Figure 1. This is a standard Prisoner Dilemma game; at equilibrium, both players exert no effort and end up with (a, a) payoffs.

When a type 1 player is matched with a type 0 player, the game can be represented by the following payoff matrix: [20]

type 1 \ type 0	0	1
0	α , α	β , δ
1	$d + qs$, β	$g + qs$, γ

Fig. 2: Type 1 v. Type 0

[18] This assumption can be replaced by a setup in which we let each pair play a repeated game and assume some learning mechanism that allows their actions to converge to Nash equilibrium actions.

[19] Therefore, when there is no subsidy, the analysis of the equilibrium actions is similar to the one described in FW.

[20] For type 1 player, we need to subtract qse^a from each term in the matrix. This does not, however, change the equilibrium strategy.

In equilibrium, the type 0 player exerts no effort, whereas the effort exerted by the type 1 player depends on the magnitude of the qs term. When $qs > a - d$, type 1 player exerts an effort and the equilibrium payoffs are $(d + qs, b)$. Otherwise, type 1 exerts no effort, which yields the equilibrium payoffs (a, a).

A game between two type 1 players can be represented by the following payoff matrix: [21]

type 1 \ type 1	0	1
0	α , α	β , $d + qs$
1	$d + qs$, β	$g + qs$, $g + qs$

Fig. 3: Type 1 v. Type 1

The solution of this game also depends upon the magnitude of the qs term:

- If $qs < b - g$, both players exert no effort at equilibrium.
- If $qs > a - d$, both players exert an effort at equilibrium.
- If $b - d < qs < a - d$, the game has two pure strategy equilbria, one equilibrium where both players exert no effort, and another equilibrium where both players exert an effort. We assume that with some (strictly)

[21] Recall that the public good and the qse^a terms have been omitted.

positive probability the players manage to coordinate on the second equilibrium.[22]

3.2 Preference Dynamics

So far, we have assumed that part of the population indeed cares about status. Intuitively, this is not surprising for most people would agree that status is an important consideration. However, justifying preferences that differ from the standard homo economicus paradigm is not trivial. The main concern of the endogenous preferences literature has been indeed to show that such preferences may still be the outcome of some preference dynamics and may survive the evolutionary process.[23]

Let $\mathrm{M}(p, p', q)$ denote the equilibrium monetary payoffs of a type p player when matched with a type p' player given q, the proportion of type 1 players in the population. Note that since the equilibrium level of effort for each type of interaction is already specified, both e^a and $\hat{e}$ are uniquely determined by q. Let $W^1(q)$ and $W^0(q)$ be the expected equilibrium payoffs of types 1 and 0, respectively:

$$W^1(q) \equiv qM(1,1,q) + (1-q)M(1,0,q)$$
$$W^0(q) \equiv qM(0,1,q) + (1-q)M(0,0,q)$$

We assume general preference dynamics, which are monotonic in the monetary payoffs. Obviously, one can conceive of population dynamics in which fitness is different from monetary payoffs. Such an assumption would probably be more realistic. However, there is no such clear and unequivocal candidate for the fitness function. We therefore choose to be conservative and to assume that fitness is simply the monetary

[22] For convenience we ignore the mixed strategy equilibrium.

[23] For example, this literature has shown the possible evolutionary stability of preferences for social status - Fershtman and Weiss (1997, 1998), fairness – Güth and Yaari (1992) and Huck and Oechssler (1999), and altruism – Bester and Güth (1998).

payoffs. Following the definition of evolutionary stability developed by Maynard Smith (1982) (see also Weibull (1995)), three types of dynamically stable populations (or preference profiles) are possible in the present model:

(a) A homogenous type 1 population is dynamically stable iff

$$W^1(1) \geq W^0(1) \text{ and } \exists m > 0 \text{ such that } \forall e < m : W^1(1-e) > W^0(1-e) \ ;^{24}$$

(b) A homogenous type 0 population is dynamically stable iff

$$W^0(0) \geq W^1(0) \text{ and } \exists m > 0 \text{ such that } \forall e < m : W^0(e) > W^1(e) \ ;^{25}$$

(c) A mixed $q \in (0,1)$ population is dynamically stable iff

$$W^1(q) = W^0(q) \text{ and } \exists m > 0 \text{ such that } \forall e < m :$$
$$W^1(q+e) < W^0(q+e) \text{ and } W^1(q-e) > W^0(q-e) .$$

We now characterize the stable preference profile and equilibrium actions given the status parameter s. Note that since type 0 players never exert an effort, the total effort $\hat{e}$, is determined by the number of type 1 players who do exert an effort.

Observation 1:

(i) When $s < b - g$, at equilibrium, both type 1 and type 0 players exert no effort in any interaction. Hence, the two types are undistinguishable in their behavior. There is no stable preference profile, and the total effort is $\hat{e} = 0$.

(ii) When $b - g < s < a - d$, the only stable preference profile is $q = 1$. All players exert an effort; and thus, $\hat{e} = N$.

(iii) When $s > a - d$, the unique evolutionary stable preference profile is $q(s) = (a-d)/s$. Type 0 players exert no effort. Type 1 players exert an effort when

[24] An equivalent condition is: $\exists m > 0$ such that $\forall q \in [1-m, 1]$ $M(1,1,q) \geq M(0,1,q)$ and $M(1,0,q) > M(0,0,q)$ if $M(1,1,q) = M(0,1,q)$.

matched with other type 1 players but they exert effort only with probability $\lambda(\sigma)$ when matched with type 0 players. $\lambda(\sigma)$ is given by:

$$(6)\quad \lambda(\sigma) = \frac{q(\sigma)(\gamma - \alpha)}{q(\sigma)\beta + (1-2q(\sigma))\alpha - (1-q(\sigma))\delta}.$$

Hence, total effort in the population is:

$$(7)\quad \hat{e}(\sigma) = q(\sigma)[q(\sigma) + (1-q(\sigma))\lambda(\sigma)]N.$$

Proof: (i) Immediate from the equilibrium behavior. (ii) First, note that type 0 players always choose $e = 0$. Hence, we focus on the equilibrium strategies of type 1 players. Given that $\beta - \gamma < \sigma < \alpha - \delta$, consider the following two possible ranges of *q*:

(1) $q \le (\beta - \gamma)/\sigma$: Type 1 players always exert no effort, and are thus indistinguishable from type 0 players. Therefore, no preference profile in this range is evolutionary stable.

(2) $q > (\beta - \gamma)/\sigma$ (note that $q \le 1 < (\alpha - \delta)/\sigma$): type 1 players, when matched with each other, sometimes exert an effort but never exert an effort when matched with type 0 players. Hence, the monetary payoff of type 1 players exceeds that of type 0 players, and *q* rises until it reaches the only stable preference profile, *q = 1*. In a stable homogenous type 1 population, every player exerts an effort.

(iii) Given that $\sigma > \alpha - \delta$, consider the following three possible ranges of *q*:

(1) $q \le (\beta - \gamma)/\sigma$: As shown in the proof of part (ii), no preference profile in this range is evolutionary stable.

(2) $(\beta - \gamma)/\sigma < q < (\alpha - \delta)/\sigma$: As shown in the proof of part (ii), the monetary payoff of type 1 players in this range exceeds that of type 0 players, and q rises.

---

[25] An equivalent condition is: $\exists m > 0$ such that $\forall q \in [0, m]$ $M(0,0,q) \ge M(1,0,q)$ and

However, contrary to the part (ii) scenario, here $(\alpha-\delta)/\sigma<1$, implying that q will continue to rise until it reaches $q=(\alpha-\delta)/\sigma$ and exits range (2). Hence, no preference profile in this range is evolutionary stable.

(3) $q>(\alpha-\delta)/\sigma$: Type 1 players always exert effort. Therefore, the monetary payoff for type 0 players exceeds that for type 1 players, and q decreases until it reaches $q=(\alpha-\delta)/\sigma$ and exits range (3). Hence, no preference profile in this range is evolutionary stable. After ruling out all other possibilities, and based upon the analysis of range (2) and range (3), we are left with $q(\sigma)=(\alpha-\delta)/\sigma$ as the unique stable preference profile.

Given the stable preference profile $q(\sigma)=(\alpha-\delta)/\sigma$, the equilibrium actions are: When two type 1 players meet, they both exert an effort. When two type 0 players meet, they both exert no effort. When a type 1 player meets a type 0 player, the type 0 player exerts no effort and the type 1 player is indifferent between exerting an effort and refraining from doing so. Hence, two outcomes are plausible: outcome (a), in which both players exert no effort, and outcome (b), in which the type 1 player exerts an effort and the type 0 player exerts no effort. Adding evolutionary stability to the equilibrium conditions, we can derive the percentage of interactions in which each one of the two outcomes occurs. Let λ denote the percentage of interactions in which the type 1 player exerts effort (i.e., outcome (b)). Evolutionary stability implies $W^1(q)=W^0(q)$ or :

$$q\gamma+(1-q)[\lambda\delta+(1-\lambda)\alpha]=q[\lambda\beta+(1-\lambda)\alpha]+(1-q)\alpha$$

Solving for λ , we obtain:

$$\lambda(\sigma)=\frac{q(\sigma)(\gamma-\alpha)}{q(\sigma)\beta+(1-2q(\sigma))\alpha-(1-q(\sigma))\delta}.$$

$M(0,1,q)>M(1,1,q)$ if $M(0,0,q)=M(1,0,q)$.

Note that $\lambda \in (0,1)$ for all values of σ in the relevant range (i.e., for all $\sigma > \alpha - \delta$).[26] Therefore, both outcomes occur with positive probabilities.

Total effort in the population, that is, the number of times that type 1 players exert an effort, is given by $\hat{e}(\sigma) = q(\sigma)[q(\sigma) + (1 - q(\sigma))\lambda(\sigma)]N$. €

3.2 The Effect of Subsidy on Effort

Assume now that given the positive externalities generated by the players' efforts, the government considers using a subsidy policy designed to encourage individuals to exert more effort. Given a direct subsidy s for exerting an effort, player i's utility function becomes:

$$(8)\quad U_i(e_i, e_j, e^a, q) = m_i + p_i\Sigma_i + se_i = \\ = \Pi_i(e_i, e_j) + p_i q\sigma(e_i - e^a) + se_i + E(\hat{e})$$

The subsidy can clearly be large enough to induce all players to exert effort. But such a policy is not necessarily optimal given the cost of raising funds. When $\sigma < \beta - \gamma$, the equilibrium behavior is that all players exert no effort; thus, $\hat{e} = 0$. In such a case, a sufficiently large subsidy could induce players to exert effort. Yet, such a case is less interesting for our current discussion because regardless of the subsidy level, the two player types remain undistinguishable; thus, no stable preference profile is possible. When $\beta - \gamma < \sigma < \alpha - \delta$, the only stable preference profile is $q = 1$. In this case, all players exert an effort and $\hat{e} = N$; therefore, there is no room for a subsidy policy. However, when $\sigma > \alpha - \delta$, the evolutionary stable preference profile is $q(\sigma) = (\alpha - \delta)/\sigma$ and total effort is $\hat{e}(\sigma) = q(\sigma)[q(\sigma) + (1 - q(\sigma))\lambda(\sigma)]N$.

[26] Also, note that $\frac{\partial \lambda}{\partial q} = \frac{(\gamma - \alpha)(\alpha - \delta)}{[q\beta + (1-2q)\alpha - (1-q)\delta]^2} > 0$.

Therefore, we choose to focus on the region of $\sigma > \alpha - \delta$ as in this region, total effort depends directly on the equilibrium distribution of preferences.

We divide our discussion into two parts. The first is the traditional short run analysis of the effects of a subsidy policy. In this part of the analysis, preferences are given at the equilibrium level $q(\sigma)$, and we show that subsidization does indeed increase total effort (and thus the level of the public good). We then proceed to the long run analysis in which the distribution of preferences may be affected by the subsidy policy. For simplicity, we assume that the policy maker is contemplating two possible policies, a no-subsidy policy and an $\hat{s}$-level subsidy policy.

3.2.1 The Effect of a Subsidy Policy in the Short Run

We first examine the effects of an $\hat{s}$-level subsidy policy under the assumption of exogenous preferences. We consider the case in which the subsidy level is insufficiently large, that is, $\hat{s} < \beta - \gamma$, such that type 0 players are not induced to exert effort. Hence, the $\hat{s}$-level subsidy policy can increase overall effort only by inducing more type 1 players to exert effort.

Observation 2: When $\sigma > \alpha - \delta$ and given the preference profile $q(\sigma)$, the use of an $\hat{s}$-level subsidy policy yields **higher** total effort in the short run.

Proof: Since $\hat{s} < \beta - \gamma$, the subsidy has no effect on the behavior of type 0 players and on the interaction between two type 1 players (in which both players exert an effort). When type 1 and type 0 are matched, recall that without subsidization, at equilibrium, type 1 players are indifferent between exerting and not exerting any effort. Adding a

subsidy $\hat{s}$ breaks this indifference.[27] Hence, type 1 will always exert an effort, and the overall effort in the economy will become $q(s)N$. Also recall that without a subsidy policy, total effort is $q(s)[q(s)+(1-q(s))\,l(s)]N$. Therefore, since $l(s)<1$, it is clear that in the short run the subsidy policy raises the overall effort level.

3.2.2 The Impact of the Subsidy Policy when Preferences are Endogenous

The subsidy policy affects the relative monetary payoffs of the two types of players. According to the general payoff monotonic dynamics defined above, a preference profile can change, implying that a subsidy policy can affect the final distribution of preferences. The following observation demonstrates that when the preference dynamics are taken into account, a subsidy policy may decrease the share of type 1 players in the population and lower total effort and the level of the public good.

Observation 3: When $s > a-d$, a subsidy $\hat{s} < b-g$ will have the following effects in the long run:

(i) The share of type 1 players in the population will decrease ; and

(ii) Total effort in the population will decrease.

Proof: (i) Recall that at the zero subsidy stable equilibrium, only $l(s)<1$ percent of those type 1 players who are matched with type 0 players exert effort (see Observation 1(iii)). Since at such an equilibrium players of type 1 who are matched with type 0 players are indifferent between exerting and not exerting an effort, the subsidy policy induces them to exert an effort whenever they are matched with type 0 players.

[27] Note that any positive subsidy, no matter how small, guarantees that type 1 players always exert effort. This discontinuity in the short run effects of a subsidy policy stems from our intentionally

Consequently, type 1's monetary payoffs decrease (since $\alpha > \delta + \hat{s}$) and type 0's monetary payoffs increase (since $\beta > \alpha$). As a result, $W^1(q) < W^0(q)$, and evolutionary dynamics drive *q* down until a new stable profile emerges. Following the logic of Observation 1(iii), the percentage of type 1 players in the new stable preference profile is: $q(\sigma, \hat{s}) = (\alpha - \delta - \hat{s})/\sigma$. Clearly:

$$q(\sigma, \hat{s}) = (\alpha - \delta - \hat{s})/\sigma < (\alpha - \delta)/\sigma = q(\sigma, 0).$$

This concludes part (i) of the proof.

(ii) At the new stable equilibrium, induced by the subsidy policy, players of type 1 exert an effort only in $\lambda(\sigma, \hat{s})$ percent of their interactions with type 0 players, where:

$$\lambda(\sigma, \hat{s}) = \frac{q(\sigma, \hat{s})(\gamma + \hat{s} - \alpha)}{q(\sigma, \hat{s})\beta + (1 - 2q(\sigma, \hat{s}))\alpha - (1 - q(\sigma, \hat{s}))(\delta + \hat{s})}.$$

We need to show that the total effort induced by the subsidy policy, $\hat{e}(\sigma, \hat{s}) = q(\sigma, \hat{s})\big(q(\sigma, \hat{s}) + (1 - q(\sigma, \hat{s}))\,\lambda(\sigma, \hat{s})\big)N$, is smaller than the total effort with a no-subsidy policy, $\hat{e}(\sigma, 0) = q(\sigma, 0)\big(q(\sigma, 0) + (1 - q(\sigma, 0))\,\lambda(\sigma, 0)\big)N$. By part (i), $q(\sigma, \hat{s}) < q(\sigma, 0)$. Hence, it is sufficient to show that:

$$(9)\quad q(\sigma, \hat{s}) + (1 - q(\sigma, \hat{s}))\,\lambda(\sigma, \hat{s}) - \big[q(\sigma, 0) + (1 - q(\sigma, 0))\,\lambda(\sigma, 0)\big] < 0.$$

Substituting the expressions derived for $q(\sigma, 0)$, $\lambda(\sigma, 0)$, $q(\sigma, \hat{s})$ and $\lambda(\sigma, \hat{s})$, we obtain, after some rearranging, that condition (9) is equivalent to:

$$(9a)\quad \frac{(\alpha - \delta - \hat{s})\big[(\beta - \gamma) - (\alpha - \delta) + \sigma\big] + \sigma(\gamma - \alpha + \hat{s})}{\sigma(\beta + \delta - 2\alpha + \sigma + \hat{s})} < \frac{(\alpha - \delta)\big[(\beta - \gamma) - (\alpha - \delta) + \sigma\big] + \sigma(\gamma - \alpha)}{\sigma(\beta + \delta - 2\alpha + \sigma)}$$

simplistic framework.

Since the denominator on the left hand side of inequality (9a) is clearly larger than the denominator on the right hand side of the inequality, we focus on the numerators. It is easy to confirm that the difference between the numerator on the left hand side of inequality (9a) and the numerator on the right hand side of the inequality is: $-[(\beta-\gamma)-(\alpha-\delta)]\cdot\hat{s}<0$. Therefore, inequality (9a) holds, and thus $\hat{e}(\sigma,\hat{s})<\hat{e}(\sigma,s=0)$ for all $\hat{s}<\beta-\gamma$.

The intuition behind these results is straightforward. With the zero subsidy benchmark, dynamic stability was obtained through type 1's discriminatory strategy. Type 1 players exert an effort whenever they play with type 1 players but when they are matched with type 0 players, they exert an effort only with some positive probability. The introduction of a subsidy causes type 1 players, in the short run, to exert an effort in all interactions, therefore allowing type 0 players to takes advantage of type 1's generosity and proliferate on her expense. The endogenous preference dynamics eventually converge to a new stable preference profile with fewer type 1 players.[28] The decline in the share of socially minded individuals, and the corresponding decrease of social incentives, more than offsets the initial rise in the monetary incentives introduced by the subsidy policy.

4. Concluding Remarks

The scope of this paper is limited as it considers only the link between public policies and preferences. But the view (or the criticism) that economists should not confine their discussion only to a world with exogenously given preferences is much

[28] The lower q induces a higher I in the new stable equilibrium. However, this secondary effect is dominated by the initial change of preferences in favor of the a-social type (type 0).

broader in scope. Market institutions and government policies may affect the evolution of values and norms of behavior as well as the evolution of preferences. This is not, however, a new approach. The effects of economic institutions on human development have been discussed ever since Alexis de Tocqueville and Karl Marx.[29]

As an additional example of the implications and generality of the approach presented at this paper, one may consider the relationship between elections and preferences. It is possible that given the voters' preference profile, a candidate is elected who implements a policy that changes the individuals' preferences in such a way that they would have been better off by electing the alternative candidate. On the other hand, one may think about reinforcement dynamics. Plausibly, the elected party will be able to implement a public policy which alters preferences in a way that would increase its attractiveness and consequently its reelection probability.

[29] For a historical perspective, see the survey by Bowles (1998).

References

Aaron, H. J., (1994), "Public policy, values, and consciousness." *Journal of Economic Perspectives* 8, 3-21.

Arrow, K. J., (1971), Political and economic evaluation of social effects and externalities. In: Intriligator, M. (Ed.), *Frontier of Quantitative Economics*, North Holland, Amsterdam.

Arrow, K. J., (1972), "Gifts and exchanges." *Philosophy and Public Affairs* 1, 343-362.

Basu, K., (1995), "Civil institution and evolution: concepts, critique and models." *Journal of Development Economics* 46, 19-33.

Becker, G., (1976), "Altruism, egoism and genetic fitness: Economics and *Journal of Economic Literature* 14, 817-826.

Bester, H., and Güth, W., (1998), "Is altruism evolutionary stable?" *Journal of Economic Behavior and Organization* 34, 193-209.

Bisin, A., and Verdier, T., (1998), "On the cultural transmission of preferences for *Journal of Public Economics* 70, 75-97.

Bowles, S., (1998), "Endogenous preferences: The cultural consequences of markets and other economic institutions". *Journal of Economic Literature* 36, 75-111.

Cavalli-Sforza, L. L., Feldman, M. W., (1981), *Cultural Transmission and Evolution: A Quantitative Approach.* Princeton: University Press.

Deci, E. (1971), "Effects of externally mediated rewards on inrinsic motivation" *Journal of Personality and Social Psychology*, 18, 105-115.

Deci, E. (1975), *Intrinsic Motivation*, Plenum Press, New York.

Dekel, E., and Scotchmer, S., (1999), "On the evolution of attitudes towards risk in *Journal of Economic Theory* 87, 125-143.

Fehr, E. Gachter, S. and Kirchsteiger, G. (1996), "Reciprocity as a contract *Econometrica*, 65, 833-860.

Fehr, E. and Gachter, S. (1998), "Reciprocity and economics: The economic implications of *Homo Reciprocans*", *European Economic Review*, 42, 845-859.

Fershtman, C., and Weiss, Y., (1997), Why do we care about what others think about us?. In: Ben Ner, A., and Putterman, L. (Eds.), *Economics, Values and Organization*. Cambridge University Press, Cambridge.

Fershtman, C., and Weiss, Y., (1998), "Social rewards, externalities and stable *Journal of Public Economics* 70, 53-74.

Frey, B. S., (1994), "How intrinsic motivation is crowded in and out." *Rationality and Society* 6, 334-352.

Frey, B. S., Oberholzer-Gee, F., and Eichenberger, R., (1996), "The old lady visits your back yard: A tale of morals and markets." *Journal of Political Economy* 104, 1297-1313.

Frey, B. S., and Oberholzer-Gee, F., (1997), "The cost of price incentives: an empirical analysis of motivation crowding-out." *American Economic Review* 87, 746-755.

Gneezy, U., and Rustichini, A., (2000a), "A fine is a price." *Journal of Legal Studies*, 29, 1-18.

Gneezy, U., and Rustichini, A., (2000b), "Pay enough or don't pay at all." *Quarterly Journal of Economics*, forthcoming.

Güth W. and Yaari, M. E. (1992), "Explaining Reciprocal Behavior in Simple Strategic Games: An Evolutionary Approach." in Witt (ed.), *Explaining Forces and Changes: Approaches to Evolutionary Economics*, Ann Arbor, University of Michigan Press.

Huck, S., and Oechssler, J., (1999), "The indirect evolutionary approach to explaining *Games and Economic Behavior* 28, 13-24.

Koçkesen, L., Ok, E. A., and Sethi, R., (2000), "Evolution of interdependent preferences in aggregative games." *Games and Economic Behavior* 31, 303-310.

Lane, R.E. (1991), *The market experience*, Cambridge: Cambridge University Press.

Maynard Smith, J., (1982), *Evolution and the Theory of Games*. Cambridge University Press, Cambridge.

Peleg, B., and Yaari, M. E., (1973), "On the existence of a consistent course of action *Review of Economic Studies* 40, 391-401.

Robson, J.A., (1996), "A biological basis for expected and non-expected utility." *Journal of Economic Theory* 68, 397-424.

Rogers, A.R., (1994), "Evoluion of time preferences by natural selection." *American Economic Review* 84, 460-481.

Solow, R.S. (1971), "Blood and Thunder" *Yale Law Journal* 80: 170-183.

Titmuss, R. M., (1970), The gift relationship" London: Allen and Unwin.

Weibull, J.W., (1995), *Evolutionary Game Theory*, MIT Press, Cambridge MA.