ISSN 1045-6333

HARVARD

JOHN M. OLIN CENTER FOR LAW, ECONOMICS, AND BUSINESS

PUBLIC POLICY WITH ENDOGENOUS PREFERENCES

Oren Bar-Gill
Chaim Fershtman

Discussion Paper No. 476

06/2004

Harvard Law School
Cambridge, MA 02138

This paper can be downloaded without charge from:

The Harvard John M. Olin Discussion Paper Series:
http://www.law.harvard.edu/programs/olin_center/

The Social Science Research Network Electronic Paper Collection:
http://papers.ssrn.com/abstract_id=######

JEL Classification: B52, H41

Public Policy with Endogenous Preferences

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

Forthcoming in the *Journal of Public Economic Theory*

Abstract: Public policy may influence norms and preferences. By altering the payoffs associated with different preferences, public policy may influence the distribution of these preferences in the population. Such interdependence between policy and preferences may limit (or enhance) the effectiveness of different policies. We demonstrate this idea with a simple model of subsidizing contributions to a public good. While the short run effect of such a subsidy will be an increase in the overall contribution, the subsidy triggers an endogenous preference change that results in a lower level of contribution to the public good, despite the explicit monetary incentives to raise that level.

Keywords: Public policy, endogenous preferences.

[*] Harvard University, The Society of Fellows, 78 Mt. Auburn St., Cambridge, MA 02138 ; And, Harvard Law School, The John M. Olin Center for Law, Economics and Business.

[**] The Eitan Berglas School of Economics, Tel Aviv University, Tel Aviv 69978, Israel.

This paper has greatly benefited from comments and suggestions by Lucian Bebchuk and seminar participants at Tel-Aviv University. We gratefully acknowledge the generous financial support provided by the John M. Olin Center for Law, Economics and Business at Harvard Law School and by the William F. Milton Fund of Harvard University.

Public Policy with Endogenous Preferences

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

© 2004 Oren Bar-Gill and Chaim Fershtman. All rights reserved.

1. Introduction

Leaders, regimes and public policies change individuals, by influencing preferences and social norms. It would be naive to think that different regimes or policies trigger only modified behavior by the citizenry, without affecting who these citizens are: their preferences, aspirations and even their dreams.[1] Such interdependence between public policies and preferences is at odds with standard economic modeling which via the exogenous preferences assumption insists on "taking individuals as they are".[2]

While neo-classical economics traditionally assumes that preferences are exogenous, economists have long recognized the malleability of individual preferences. Half a century ago Harsanyi (1953-1954) wrote that: *"the Economic Problem of a community… includes also the question of how these scarce resources should be divided between productive operations for satisfying people's actual wants and measures for changing these wants."* The potential effect of public policy on individual preferences

[*] Harvard University, The Society of Fellows, 78 Mt. Auburn St., Cambridge, MA 02138 ; And, Harvard Law School, The John M. Olin Center for Law, Economics and Business.

[**] The Eitan Berglas School of Economics, Tel Aviv University, Tel Aviv 69978, Israel.

This paper has greatly benefited from comments and suggestions by Lucian Bebchuk and seminar participants at Tel-Aviv University. We gratefully acknowledge the generous financial support provided by the John M. Olin Center for Law, Economics and Business at Harvard Law School and by the William F. Milton Fund of Harvard University.

[1] See Aaron (1994), Bowles (1998) and Marschak (1978). John Stuart Mill argued *"government itself should be evaluated in large measure by its effects on the character of the citizenry."* (cited in Sunstein 1997, p. 20)

[2] For a clear early statement of this approach see Stigler and Becker (1977) and Becker (1976): "… all human behavior can be viewed as involving participants who maximize their utility from a stable set of preferences…."

has also been recognized. Referring to environmental policy questions, Sen (1995) writes:

> *"There are plenty of "social choice problems" in all this, but in analyzing them, we have to go beyond looking only for the best reflection of given preferences, or the most acceptable procedures for choices based on those preferences. We need to depart both from the assumption of given preferences (as in traditional social choice theory) and from the presumption that people are narrowly self-interested homo economicus (as in traditional public choice theory)."*

Nevertheless, there has been no attempt to formally model the interdependence between public policy and individual preferences and its implication on the effectiveness of different public policies.[3] Building on the recent literature on the endogenous formation of preferences this paper develops such a model. We demonstrate how public policy, by altering the payoffs associated with different preferences, affects the dynamics of preference formation which in turn influence the efficacy of the public policy.

In recent years there is a growing literature that studies the endogenous formation of preferences. Preferences may evolve as a result of cultural transmission by which a socialization process transmits preferences across generations or through an imitation process by which individuals imitate other more "successful" individuals.[4] Such endogenous preference dynamics introduces a direct link between public policy and the

[3] An important exception is a series of recent papers studying the interaction between legal policy and preferences. Huck (1998) analyzes the effects of the cost, effectiveness and outcome of a legal monitoring process on the evolution of remorse, and shows how legal institutions can be designed to encourage mutual trust. Similarly, Bohnet et al. (2001) show how a small probability of contract enforcement can "crowd-in" trustworthiness, or preferences for honesty, and thus lead to a higher probability of performance. And, Bar-Gill and Fershtman (2004) study the effects of the legal remedy for breach of contract on the evolution of fairness norms and preferences among contracting parties.

[4] Cultural transmission dynamics have been studied by Cavalli-Sforza and Feldman (1973, 1981), Boyd and Richerson (1985) and Bisin and Verdier (1998, 2001). For evolutionary models that endogenize preferences - see Basu (1995), Bester and Güth (1998), Dekel and Scotchmer (1999), Fershtman and Weiss (1997,1998), Guth and Yaari (1992), Huck and Oechssler (1999), Koçkesen, Ok and Sethi (2000a,b), Possajennikov (2000), Robson (1996), and Rogers (1994).

formation of preferences. Public policy changes the outcomes of the market interactions and thus affects the evolution of future preference profiles.[5] Moreover, the new distribution of preferences, or the altered social norm, may now affect the players' behavior and their reaction to the implemented public policy. Such interdependence between policies and preferences may limit (or enhance) the efficacy of public policies and thus influence the design of optimal policy.

We demonstrate this idea by considering a simple public good model in which contributions to the public good are encouraged by direct monetary subsidization. Assuming endogenous preference dynamics we consider the effect of such government subsidization on the formation of preferences and consequently on the long run level of the public good.

Individuals in this model are pair-wise matched and play a strategic game. The players' actions affect their direct payoffs and the aggregate action contributes to 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, all players do not necessarily share this concern for social status. We do not impose any preference profile; rather, we assume a selection process that determines the profile endogenously.

We then consider a subsidy policy aimed at promoting contributions to the public good. A short run analysis, in which preferences are exogenously given, indicates that such a subsidy indeed increases the equilibrium level of the public good. In the long run,

[5] Such interdependence introduces also conceptual difficulties in studying, or defining, optimal public policies. The standard modeling approach is to define optimal policy with respect to an optimality criterion that ranks market outcomes given an exogenously specified profile of preferences. However, when public policy affects the evolution of preferences, the above selection procedure is no longer valid.

however, the subsidy policy induces a shift in the distribution of preferences, reducing the social incentives as well as the proportion of the population that cares about them. Consequently, in our model 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.

Our emphasis is on social concerns, which are arguably more sensitive to the type of preference formation process that we consider. The importance of social rewards in providing incentives or compensation for individuals who perform activities with positive externalities was already suggested by Arrow (1971). We emphasize the possible limits of standard monetary incentives in inducing activities that also provides social benefits. We demonstrate how the use of monetary incentives may trigger an endogenous preference shift. This shift implies not only lower social incentives, but also a smaller proportion of individuals who care about social rewards.

The above result resembles a well-known argument in social psychology. In a controversial book, Titmuss (1970) argued that allowing payments for blood donations would result in a lower level (and even in a lower quality) of donation.[6] This hypothesis suggests that in some circumstances, monetary rewards crowd out intrinsic motivation like civic duty or altruism. The "crowding out" problem has been studied by cognitive social psychologists[7] as well as by economists.[8] While our model predicts similar outcomes, the underlying phenomenon is quite different. The emphasis in the

[6] See also Solow (1971) and Arrow (1972) for a critical review of this argument.

[7] For a survey of this literature see Deci (1975) and Lane (1991).

[8] See, e.g., Frey (1994), Frey et al. (1996) and Frey and Oberholzer-Gee (1997) for a detailed examination of the crowding-out effect. See also Huck (1998) and Bohnet et al. (2001). For recent experimental studies on the effectiveness of monetary incentives and the crowding in or out effects - see Fehr, Gachter and Kirchsteiger (1996), Fehr and Gachter (1998), and Gneezy and Rustichini (2000a,b).

psychological and experimental economics literature is on the effect of monetary incentives on people's perception of the social rewards from "honorable" activities. That is, if a price is placed on a blood donation, then donating blood may no longer be considered a noble act or 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.[9] Such a preference shift may affect a real change in the relative importance of intrinsic motivation and extrinsic monetary rewards.

2. Subsidizing a Public Good: Incentives and Preference Formation

We consider a population that interacts strategically. The action that each player chooses, besides affecting his direct private payoff, contributes to the accumulation of a public good and also determines the player's social status. Preferences, specifically preferences for social status, may change over time depending on the relative success of different types of individuals.

2.1 The Market Interaction

We follow Fershtman and Weiss (1997) and 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, $e_i \in \{0,1\}$. Let $\Pi_i(e_i, e_j)$ be the direct monetary payoff of player *i*, who is matched with player *j*. The values of $\Pi_i(.,.)$ are given in the following payoff matrix:

[9] The change in preferences may indirectly influence the magnitude of the social rewards associated with contribution to the public good. See Section 2 (specifically equation (4)) below.

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

Fig. 1: The Payoff Matrix

where $\beta > \gamma > \alpha > \delta$. We further assume that $\alpha - \delta > \beta - \gamma$.

In addition to the direct payoff, the players' overall efforts contribute to a public good that they all commonly enjoy. Let $\hat{e}$ be the total amount of effort in the population and let $E(\hat{e})$ be a public good term such that $E(\cdot), E'(\cdot) > 0$. 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. Player *i*'s overall payoff is: [10]

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

We further assume that e_i also determines the individual's social status.[11] 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

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

[11] As is conventional in the endogenous preferences literature, we adopt a specific social preference – a preference for status. Our main point, however, is general: if a different social preference were assumed, a

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 to the public good. Letting the average effort e^a, $e^a = \hat{e}/N$, represent the social norm;[12] we assume that status (positive or negative) is conferred upon an individual according to her performance relative to the social norm.

We further assume that only socially minded individuals can confer status on others. Under such an assumption the magnitude of the social incentives depends on the distribution of preferences in the population.[13] Specifically, the social component in individual *i*'s utility function is

$$(4)\quad \Sigma_i \equiv q\sigma(e_i - e^a),$$

where σ 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* -

different model could be constructed where public policy affects the distribution of preferences in the population.

[12] The social norm is also endogenously determined. If a group of individuals does not obey the norm, this will change the norm itself.

[13] The notion is that an individual cares about his relative position, or status, because he cares about other individuals' opinion of him. In a society where individuals do not appreciate a certain trait or a certain behavior, possessing this trait or adopting this behavior would not be as important. Thus, when allowing for social preferences, the distribution of social and asocial types should affect the individual's utility function.

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

We assume that the players' types are fully observable. 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.

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

When a type 1 player is matched with a type 0 player, the game can be represented by the following payoff matrix: [14]

		type 0	
		0	1
type 1	0	α , α	β , δ
	1	$\delta + q\sigma$, β	$\gamma + q\sigma$, γ

Fig. 2: Type 1 v. Type 0

In equilibrium, the type 0 player exerts no effort, whereas the effort exerted by the type 1 player depends on the magnitude of the $q\sigma$ term. When $q\sigma > \alpha - \delta$, the type 1 player exerts effort and the equilibrium payoffs are $(\delta + q\sigma, \beta)$. Otherwise, the type 1 player exerts no effort, and the equilibrium payoffs are (α, α).

A game between two type 1 players can be represented by the following payoff matrix: [15]

		type 1	
		0	1
type 1	0	α , α	β , $\delta + q\sigma$
	1	$\delta + q\sigma$, β	$\gamma + q\sigma$, $\gamma + q\sigma$

Fig. 3: Type 1 v. Type 1

[14] For type 1 player, we need to subtrac ^{a} from each te he matrix. This does not, however, change the equilibrium strategy

The solution of this game also depends on the magnitude of the $q\sigma$ term: If $q\sigma < \beta - \gamma$, both players exert no effort at equilibrium. If $q\sigma > \alpha - \delta$, both players exert effort at equilibrium. If $\beta - \gamma < q\sigma < \alpha - \delta$, the game has two pure strategy equilbria, one equilibrium where both players exert no effort, and another equilibrium where both players exert effort. We assume that with some (strictly) positive probability the players manage to coordinate on the second equilibrium.[16]

2.2 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.[17] However, justifying preferences that differ from the standard *homo economicus* paradigm is not trivial.[18]

[15] Recall that the public good and the $q\sigma e^a$ terms have been omitted.

[16] We ignore the mixed strategy equilibrium. Our results would not change were we to focus on the mixed strategy equilibrium instead (for our results to hold, all that is required is that when $\beta - \gamma < q\sigma < \alpha - \delta$ the expected payoff of a type 1 player is greater than α; this requirement is satisfied in the mixed strategy equilibrium).

[17] Adam Smith (1776) wrote "Honour makes a great part of the reward of all honourable professions." (*The Wealth of Nations*, Book 1, ch. X, part1). Max Weber (1922) was the first to introduce social status as an important source of power. He defined status as "an effective claim to social esteem in terms of negative or positive privileges" [reprinted 1978, p.305].

[18] The main concern of the endogenous preferences literature has been to show that such preferences may still be the outcome of some preference dynamics and may survive the evolutionary process. For a derivation of the conditions under which "standard" preferences survive, see, e.g., Guth and Peleg (2001).

Let *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 payoff. We therefore choose to be conservative and to assume that fitness is simply the monetary payoff.[19] Following the definition of evolutionary stability developed by Maynard Smith (1982) (see also Weibull 1995), a homogenous population of type *p* players is evolutionarily stable (i.e. an ESS) *if and only if* for any possible type $p' \neq p$ either – (a) a type p' player earns a lower payoff against a type *p* player as compared to the payoff earned by a type *p* player when matched against another type *p* player; or (b) if both type *p* and type p' players earns the same payoff when matched against a type *p* player, then a type *p* player earns a higher payoff against a type p' player as compared to the payoff earned by a type p' player when matched against another type p' player. A mixed population, $q^* \in (0,1)$, is dynamically stable if the two types earn the same

[19] Here, as in other endogenous preferences models, the discussion regarding the appropriate assumptions about preferences is replaced by a discussion about the appropriate fitness criterion. This discussion is beyond the scope of this paper. Non-monetary (specifically social) factors may clearly enter into the fitness function. However, there is no clear and unequivocal candidate for a non-monetary fitness criterion. Moreover, we chose a monetary fitness function to show that even with such a conservative fitness function (that minimizes the deviation from the homo-economicus paradigm) at equilibrium individuals care about status. Clearly, our main point regarding the effect of policy on the distribution of preferences does not depend on our choice of a monetary fitness function, though a different fitness function would likely entail a different manifestation of this point.

expected payoffs given q^*, but whenever $q > q^*$ type 0 gets a higher payoff than type 1 and whenever $q < q^*$ type 1 gets the higher payoff.

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

<u>**Proposition 1**</u>:

(i) When $\sigma < \beta - \gamma$, at equilibrium, both type 1 and type 0 players exert no effort in any interaction. Hence, the two types are undistinguishable in their behavior. As a result any preference profile $q \in [0,1]$ is neutrally stable (i.e., an NSS; see Weibull 1995 for a formal definition) and the total effort is $\hat{e} = 0$.

(ii) When $\beta - \gamma < \sigma < \alpha - \delta$, the only stable preference profile is q = 1. All players exert effort; and thus, $\hat{e} = N$.

(iii) When $\sigma > \alpha - \delta$, the unique evolutionary stable preference profile is $q(\sigma) = (\alpha - \delta)/\sigma$. Type 0 players exert no effort. Type 1 players always exert effort when 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:

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

Proof: See Appendix.

The intuition for this result is as follows:

(i) With weak status concerns the behavior of the two types of players are indistinguishable.

(ii) With intermediate status concerns, type 1 players, when matched with each other, sometimes exert effort, but they never exert 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, $q = 1$.

(iii) With strong status concerns, the evolutionary stable population necessarily consists of both type 1 and type 0 players, where the stable distribution of preferences is determined as follows: If the proportion of type 1 players is too large, these players will always exert effort (even when matched with type 0 players), and therefore will earn lower payoffs than type 0 players, pushing down the proportion of type 1 players. On the other hand, if the proportion of type 1 players is too low, these players will sometimes exert effort when matched with each other, but will never exert effort when matched with type 0 players. As a result, type 1 players will earn higher payoffs than type 0 players, pushing up the proportion of type 1 players. At equilibrium, type 1 players, when matched with each other, always exert effort, but they also sometimes exert effort when matched with type 0 players. Evolutionary stability, implying that the average payoff of type 1 players equals the average payoff of type 0 players, determines the probability with which type 1 players will exert effort when matched with type 0 players.[20] This probability, together with the proportion of type 1 players in the population also determines total effort.

2.3 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 effort, player i's utility function becomes:

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

When $\sigma < \beta - \gamma$, in equilibrium all players exert no effort and $\hat{e} = 0$. 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 with a sufficiently large subsidy, the two types remain undistinguishable. When $\beta - \gamma < \sigma < \alpha - \delta$, the only stable preference profile is q = 1. In this case, all players exert effort and $\hat{e} = N$; therefore, there is no room for a subsidy policy. When $\sigma > \alpha - \delta$, the evolutionary stable preference profile is $q(\sigma) = (\alpha - \delta)/\sigma$ and total effort is $\hat{e}(\sigma) = q(\sigma)[q(\sigma) + (1 - q(\sigma))\lambda(\sigma)]N$ (see proposition 1(iii)). Therefore, we choose to focus on the $\sigma > \alpha - \delta$ region, since in this region total effort depends directly on the equilibrium distribution of preferences.

[20] This extended notion of evolutionary stability, where λ is set to attain a rest point of the dynamic process, is not necessary for the analysis. Alternatively, we could arbitrarily fix λ, and have $q(\sigma) = (\alpha - \delta)/\sigma$ not as a rest point, in which expected payoffs are identical, but rather as convergence point, such that every deviation will cause the system to converge back to $q(\sigma) = (\alpha - \delta)/\sigma$. It can be

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 the preference profile is given at the equilibrium level of $q(\sigma)$, and we show that subsidization does indeed increase total effort and consequently 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 a $\hat{s}$-level subsidy policy.

2.3.1 The Effect of a Subsidy Policy in the Short Run

We first examine the short run effect of a $\hat{s}$-level subsidy assuming a given profile of preferences. We restrict our analysis to low-level subsidies i.e., $\hat{s} < \beta - \gamma$, such that the subsidy is insufficient to induce type 0 players to exert effort.[21] Hence, the $\hat{s}$-level subsidy policy can increase overall effort only by inducing more type 1 players to exert effort.

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

The intuition for this result, whose detailed proof is omitted, is as follows: Since $\hat{s} < \beta - \gamma$, the subsidy has no effect on the behavior of type 0 players. Nor does the subsidy affect the interaction between two type 1 players (in which both players exert effort). When type 1 and type 0 players are matched, recall that without subsidization, at

readily verified that all the results continue to hold under this alternative notion of asymptotic stability (for any constant λ).

equilibrium, type 1 players are indifferent between exerting and not exerting effort, and will exert effort with a certain probability, λ. Adding a subsidy $\hat{s}$ breaks this indifference. Hence, type 1 will always exert effort, and the overall effort in the economy will increase.

2.3.2 Subsidy Policy with Endogenous Preferences

The subsidy policy affects the relative monetary payoffs of different types of players. Hence, the general payoff monotonic preference dynamics that we described imply that the subsidy policy may affect the final distribution of preferences. The following proposition demonstrates that when preference dynamics are taken into account, a subsidy policy may decrease the share of type 1 players in the population and consequently lower total effort and the level of the public good.

Proposition 3: When $\sigma > \alpha - \delta$, a subsidy $\hat{s} < \beta - \gamma$ will have the following effects:

(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 $\lambda(\sigma) < 1$ percent of those type 1 players who are matched with type 0 players exert effort (see Proposition 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 effort, the subsidy policy induces them to exert effort whenever they are matched with type 0 players. Consequently, type 1's monetary payoff decreases (since $\alpha > \delta + \hat{s}$; recall that $\hat{s} < \beta - \gamma$

[21] A sufficiently large subsidy can clearly induce all players to exert effort. We assume, however, that the

and $\alpha - \delta > \beta - \gamma$) and type 0's monetary payoff increases (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 Proposition 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:

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

(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:

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

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

costs (real or political) of funding a subsidy $s > \hat{s}$ render such a subsidy unattractive if not unfeasible.

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: $-\left[(\beta-\gamma)-(\alpha-\delta)\right]\cdot\hat{s}<0$. Therefore, inequality (9a) holds, and thus $\hat{e}(\sigma,\hat{s})<\hat{e}(\sigma,s=0)$ for all $\hat{s}<\beta-\gamma$. □

The intuition for this result is as follows: In the zero subsidy benchmark, dynamic stability was obtained through type 1's discriminatory strategy. Type 1 players exert effort whenever they are matched with other type 1 players, but only exert effort with some positive probability when they are matched with type 0 players. 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. Therefore, the endogenous preference dynamics eventually converge to a new stable preference profile with fewer type 1 players.[22] 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.

The above effect of the subsidy policy may be alternatively stated in terms of a tax policy.

Corollary: When preferences are determined endogenously, a tax policy may be effective in promoting contributions to a public good, and may thus increase the overall level of the public good (independent of any direct spending of the tax revenues on the public good).

3. Conclusion

The claim that market institutions and government policies may affect the evolution of values and norms of behavior as well as the evolution of preferences has been discussed ever since Alexis de Tocqueville and Karl Marx.[23] The main goal of such claims has been to criticize mainstream economics and its underlying premise of exogenous preferences. There has been no attempt to formally model the implications of this critique for the optimal design of public policy. Moreover, little has been said regarding the precise mechanism through which public policy may affect norms and preferences.

This paper has combined insights from the growing literature on dynamic preference formation into a model of public policy design and has provided an example of a possible formalization of the interdependence between public policy, preferences and norms. Using this formalization, the analysis has demonstrated the possible counterintuitive conclusions that follow from this interdependence.

The study of optimal policy under endogenous preferences clearly reaches beyond the question of subsidizing the accumulation of public goods that was examined in this paper. For instance, the political economy models that study the relationship between elections, voters' preferences and public policy can and should be enriched by an explicit account of the dynamic interaction between policy and preferences.

[22] The lower q induces a higher λ 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).

[23] 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.

Bar-Gill, O., and Fershtman, C., (2004), “Law and preferences.” *Journal of Law, Economics and Organization*, forthcoming.

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

Becker, G., (1976), “*The Economic Approach to Human Behavior*”, Chicago: University of Chicago Press.

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 social status.” *Journal of Public Economics* 70, 75-97.

Bisin, A., and Verdier, T., (2001), “The Economics of Cultural Transmission and the Evolution of Preferences” *Journal of Economic Theory*, 97(1), pp.298-319.

Bohnet, I., Frey, B. S. and Huck, S. (2001), “More Order With Less Law: On Contract Enforcement, Trust and Crowding,” *American Political Science Review*, 95, 131-144.

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

Boyd, R. and P. Richerson (1985) *Culture and the Evolutionary Process*, Chicago: Chicago University Press.

Cavalli-Sforza, L. L., Feldman, M. W., (1973), “Cultural versus Biological Inheritance: Pheno-type Transmission from Parent to Children “ *American Journal of Human Genetics*, 25, pp. 618-37.

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

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

Dekel, E., and Scotchmer, S., (1999), “On the evolution of attitudes towards risk in winner-take-all games.” *Journal of Economic Theory* 87, 125-143.

Fehr, E. Gachter, S. and Kirchsteiger, G. (1996), “Reciprocity as a contract enforcement device”, *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 E. Kalai (1997) "Unobserved Delegation" *International Economic Review* 1997 (November), Vol. 38, No. 4 pp. 763-774.

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 preferences.” *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 Peleg, B., (2001), “When will payoff maximization survive? an indirect evolutionary analysis.” *Journal of Evolutionary Economics* 11, 479-499.

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.

Harsanyi, J. C. (1953-54), “Welfare Economics of Variable Tastes,” *Review of Economic Studies*, 21, 204-213.

Huck, S. (1998), “Trust, Treason, and Trials: An Example of How the Evolution of Preferences Can Be Driven by Legal Institutions,” *Journal of Law, Economics and Organization*, 14, 44-60.

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

Koçkesen, L., Ok, E. A., and Sethi, R., (2000a), “The Strategic advantage of Negatively Independent Preferences” *Journal of Economic Theory* 92, 274-299.

Koçkesen, L., Ok, E. A., and Sethi, R., (2000b), “Evolution of Interdependent Preferences in Aggregative Games” *Games and Economic Behavior*, 31, pp.303-310.

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

Marschak, T. A. (1978), “On the Study of Taste Changing Policies,” *American Economic Review*, 68(2), 386-391.

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

Possajennikov, A. (2000), “On the Evolutionary Stability of Altruistic and Spiteful Preferences” Journal of Economic Behavior and Organization” 42(1), pp.125-129.

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

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

Sen, A. K. (1995), “Rationality and Social Choice”, *American Economic Review*, 85, 1-24.

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

Smith, A. (1776), *The Wealth of Nations*. Reprint, Modern Library, New-York, 1937.

Stigler, G. J. and G. S. Becker (1977) “De Gustibus Non Est Disputandum” *American Economic Review*, 67, pp. 76-90.

Sunstein, C. R. (1997), *Free Markets and Social Justice* (Oxford University Press, New York, NY).

Titmuss, R. M., (1970), *The Gift Relationship* (Allen and Unwin, London, UK).

Weber, M. (1922), *Economy and Society,* Reprinted: University of California Press, Berkeley, 1978.

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

Appendix

Proof of Proposition 1:

(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 never exert 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 effort, but they never exert 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 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. 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 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 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 :

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

Solving for λ, we obtain:

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

Note that $\lambda \in (0,1)$ for all values of σ in the relevant range (i.e., for all $\sigma > \alpha - \delta$).[24] Therefore, both outcomes occur with positive probabilities.

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

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