Centre de Referència en Economia Analítica

Barcelona Economics Working Paper Series

Working Paper nº 102

A Theory of Endogenous Sentiments

Joan Esteban and Laurence Kranich

January, 2003

A Theory of Endogenous Sentiments

Joan Esteban* and Laurence Kranich†

Preliminary.
Please do not quote without permission of the authors.

Barcelona Economics WP nº 102

January, 2003

Abstract

We present a model in which each agent's sentiments toward others are determined endogenously on the basis of how they behave relative to a standard of appropriate behavior. As sentiments change, so too does the optimal behavior of each individual, which in turn affects other agents' sentiments toward them. We focus on fixed points of this reciprocal adjustment process. To demonstrate the potential use and implications of such a model, we present an extended example involving team production. We then consider various standards of behavior, and we examine stationary patterns of behavior and sentiments under each.

JEL CLassification: D50; D64; D31

Keywords: Endogenous altruism; Group formation; Social contract

*Institut d'Analisi Economica, CSIC, and Universitat Pompeu Fabra. Mailing address: Campus de la UAB, 08193 Bellaterra-Barcelona, Spain. Email: Joan.Esteban@uab.es

†Department of Economics, University at Albany, SUNY. Mailing address: 1400 Washington Ave., Albany, NY 12222. Email: L.Kranich@albany.edu

1. Introduction

Various behaviors are difficult to explain within the standard, rational choice framework involving purely selfish individuals. Benevolent acts, such as gift-giving and volunteerism, as well as malevolent acts, such as vandalism or revenge, impose a cost on the perpetrator with no apparent benefit. In attempting to exhibit or model such behavior, a common procedure has been to assume that agents are affected by the consequences of an act or by the act itself.[1] For example, if agents are altruistic, they might engage in behavior to enhance the well-being of others. Or if they are concerned about their relative position in society, then they might engage in harmful behavior in order to worsen the position of others and thereby improve their own relative standing. Alternatively, agents might derive satisfaction (e.g., a *warm glow*[2]) simply from behaving generously, regardless of the state or even identity of the beneficiary. Etc.

By tailoring the individual to the circumstances in this way (for example, by specifying preferences accordingly), one can explain a wide range of behavior. However, if the same individual were then placed in different circumstances, the model would lose most of its explanatory power. Indeed, as Matthew Rabin points out in [14], and as is amply demonstrated in the experimental literature, actual behavior is quite complex; the same individual might be benevolent toward some agents and malevolent toward others or even benevolent or malevolent toward the same agent at different times. In particular, Rabin argues that "people like to help those who are helping them, and to hurt those who are hurting them." Similarly, Ernst Fehr and Simon Gächer [9] conclude that such reciprocity is a robust phenomenon in experimentation, even though the cause is still under debate.

Several theoretical models have been proposed to capture reciprocal behavior. For example, Rabin's paper contains a 2-agent model in which it is possible to infer motive from an opponent's actions. Then, given the same material payoff, agents might be affected differently, and hence behave differently, depending on what they

[1] It is also possible to focus on the indirect (selfish) benefits (e.g., tax advantage, reciprocation, or recognition and social status). For example, parents might transfer to children at one stage in order to engender reciprocal transfers at a later stage. (Cf., Becker, Bergstrom, Bernheim) Similar explanations have been afforded for transfers outside the family. Also, as an insurance mechanism, one might support social structures that benefit others if one might one day find oneself in need of assistance.

[2] That is, the sheer joy of giving without regard to the welfare of the recipient. (See Andreoni [1].)

perceive to be their opponent's motive.[3] In particular, they may wish to punish those they perceive as acting against their interests and to reward those they perceive as acting in favor of their interests, regardless of the material outcome.

Similarly, David Levine [13] presents a model in which agents' preferences consist of a linear combination of their own monetary payoff and the material payoffs of other agents. Moreover, the extent of their concern for others, i.e., the weight (positive or negative) on the opponent's income, is private information. Agents then modify such weights on the basis of their perception of other agents' concern for them, increasing the weights assigned to altruistic agents and decreasing the weights of those who are spiteful.[4]

More generally, John Geanakoplos, David Pearce and Ennio Stacchetti [11] introduced the concept of a *psychological game* in which a player's overall payoff depends not only on their material payoff but on their beliefs as well.[5]

Such models, by allowing the same individual to exhibit different and possibly conflicting sentiments, significantly enhance the explanatory power of the theory. However, they still present several inconsistencies with observed phenomena. First, as presently formulated, such reciprocal models generally pertain to two agent settings and thus, by definition, exclude third party effects.[6] However, we would argue that our impressions of others, and hence sentiments toward them, are often affected by their treatment of third parties. For example, witnessing selfless or heroic acts might affect our esteem for the actors even if they have no impact on one's own material well-being. Similarly, acts of cruelty are likely to affect one's view of the perpetrator even if one is not the target or victim of aggression.

But even if the extant reciprocal models could be extended to include additional agents, the focus of the analysis would nevertheless be on the effect of other agents' actions *on the observer*.[7] Hence, they may be able to explain quid pro quo

[3] See also Oded Stark and Ita Falk [16].

[4] Other references include Bolton and Ockenfels [2], Bowles and Gintis [4],Charness and Haruvy [5], Dufwenberg and Kirchsteiger [6] Falk and Fischbacher [8], and Fehr and Schmidt [10], among others.

[5] On this, see also Itzhak Gilboa and David Schmeidler [12].

[6] Most experiments in this area consider diadic, often ultimatum, games.

[7] There is some question as to whether such models can be extended. For example, in Rabin's model, agent i infers the motive of agent j by noting the effect of j's behavior on i's payoff relative to that which j could have acheived. However, with additional agents, the causal link is broken; it may no longer be possible to associate the outcome with j's behavior alone, rather it is jointly determined by the actions of all others. Thus, the ability to the infer motive of a particular

behavior among agents who meet face to face, but they are less adept at explaining anonymous acts – such as international development or famine relief efforts – where reciprocation is unlikely to play a role. In light of such efforts, it would seem that people often help those who are least likely to help them in return.

Here, we present an alternative model in which agents' sentiments toward others are determined endogenously on the basis of their behavior, with no attempt to infer other agents' motives or sentiments. While the model allows for third party effects, it is primarily motivated by two general observations. First, it is often the case that we judge the behavior of others relative to their circumstances. Thus, for example, if the same two individuals were to engage in the same benevolent act (possibly with the same motive) – say a charitable contribution of \$100 – but one were rich and the other poor, then our appreciation of the two might be quite different. Conversely, if both were to fail to act benevolently, our degree of disappoint in the two might differ as well. Alternatively, our expectations for the behavior by a healthy individual might differ considerably from those for one who is ill.

Our second observation is that, in directing our sympathies, we often take into consideration the role of the individual in determining his or her circumstances, allowing scope for individual responsibility. For instance, we might have much less sympathy for someone who squandered considerable personal wealth than for someone else who is misfortunate *through no fault of their own.* Or, we might feel differently toward victims of a famine that results from political versus natural causes.

These observations indicate different treatments in a formal model of behavior. The former requires that we focus on a fixed point or steady state of a reciprocal process in which sentiments change in response to behavior and vice versa. The latter requires that we specify how agents evaluate appropriate behavior. We incorporate both treatments in our model of endogenous sentiments.

Rather than present a general model, we instead demonstrate the method by means of an extensive example involving team production. That is, a group of agents jointly contribute labor to the production of a single consumption good, and each agent must decide how much labor to supply. We take the distributive shares of output to be given. Thus, the circumstances of the agents consist of their productivities as well as their output shares. In other contexts, the nature of agents' "circumstances" might differ as might the nature of their contributions.

individual is obfuscated. As a simple example, agent j's actual motive might be to affect a third party, k, and the effect on i might be purely incidental.

The manner in which the methodology might be applied in other contexts will be obvious. But in any event, the focus of attention should be on the method of preference formation rather than on the specific example. Our objective in this paper is simply to motivate the approach by demonstrating its potential explanatory power.

In brief, as in the aforementioned work, we present a model in which agents have extended preferences defined over the welfare of others, where the welfare weights are determined endogenously. However, unlike the antecedents, we assume each agent formulates a *standard of appropriate behavior* for every other agent in light of their circumstances. Then agent i modifies its concern for agent j, i.e., its welfare weight, on the basis of the deviation between j's actual behavior and that which i thinks is appropriate: if j's behavior surpasses i's standard, then i increases the weight it attaches to j's welfare, and if i is disappointed by j's behavior, it decreases its weight. In turn, as its weights change, i's (actual) behavior will change as well. We then consider stationary patterns of utility interdependence, that is, stationary patterns of mutual concern.

To completely specify the model, it is necessary that we state what is meant by "appropriate behavior." While there may be many such formulations, here we offer three possibilities. First, we consider the case in which agents take the mean behavior as the societal norm and judge other agents' actions accordingly.[8] In the second case, to determine what agent i thinks is appropriate behavior for agent j, we consider what i would do if it were in j's shoes, that is, if it faced similar circumstances. And finally, we suppose that each agent has a preconceived notion of what constitutes a *fair labor contribution* in light of the circumstances, i.e., the agents' productivities and their output shares. In each of the three cases, our objective is to demonstrate how sentiments might be reciprocally affected and what possible stationary patterns might emerge.

The paper is organized as follows. In the next section, we present the team production problem which we use to develop the basic structure of the model of endogenous sentiments. We emphasize that this is primarily an organizational device; the methodology can be greatly extended and applied outside of this particulat context. Next, in Sections 3-5 we consider, in turn, three possible standards of behavior. Finally, Section 6 contains concluding remarks and a blueprint for future work.

[8]For example, if the norm is to work 40 hours per week, then one might measure agents' industry relative to this standard.

2. The model

We consider an n-agent team production problem in which agents contribute heterogeneous labor to the production of a single output. Let $N = \{1, ..., n\}$ denote the set of agents. For simplicity, we assume the technology is described by the Cobb-Douglas production function

$$Y = f(\mathbf{L}) = \prod_{i=1}^{n} L_i^{\beta_i}, \text{ where } \beta_i \geq 0 \text{ for all } i \in N, \text{ and } \sum_i \beta_i = 1,$$

where $\mathbf{L} = (L_1, ..., L_n)$ denotes the vector of labor inputs and β_i measures the productivity of agent i.

Output is allocated among the agents according to fixed distributive shares in which i receives $\theta_i Y = c_i$.[9] Thus, agent i's "circumstances" can be described by (β_i, θ_i). We denote $(\boldsymbol{\beta}, \boldsymbol{\theta}) = ((\beta_i, \theta_i))_{i \in N}$.

Each agent has direct, or egoistic, preferences for consumption and labor which are represented by a utility function of the form $u_i(c_i, L_i)$. In addition to its direct preferences, it also has an extended, or social, utility function of the form

$$U_i(\mathbf{c}, \mathbf{L}) = \sum_j \alpha_{ij} u_j(c_j, L_j), \tag{2.1}$$

where $\alpha_{ii} \equiv 1$ and $\alpha_{ij} \leq 1$ for all $j \in N$. Let $\alpha_i = (\alpha_{i1}, ..., \alpha_{in})$ and $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_n)$.

Substituting for c_j in (2.1), we write the extended utility function as

$$U_i(\mathbf{L}) = \sum_j \alpha_{ij} u_j(\theta_j \prod_{i=1}^{n} L_i^{\beta_i}, L_j). \tag{2.2}$$

The objective of the paper is to study the endogenous determination of the coefficients α_{ij}. Briefly, suppose each agent were to choose its own labor supply. Other agents would then judge the appropriateness of such decisions and modify the weight they attach to the utilities of others, i.e., i would modify α_{ij} on the basis of j's choice of L_j This, in turn, will affect i's own behavior, which would then affect α_{ki} and hence L_k, for $k \neq i$. (For example, agents might work harder for the common good were they to share a greater sense of concern for others.[10])

[9] For example, consider a partnership in which the profit shares are negotiated at the outset.

[10] We test this hypothesis in Section 4, below.

A stationary outcome of the model consists of a vector of sentiments $\boldsymbol{\alpha}$ and labor supplies $\mathbf{L}$ such that no further modification occurs. The focus of the paper is on such outcomes.

Formally, let L_{ij} denote agent i's perception/evaluation of what appropriate behavior for j – in contrast to j's actual labor supply, L_j. Then, as indicated above, we envision the coefficients $\boldsymbol{\alpha}$ evolving according to a dynamic process. Treating time as a discrete variable, we assume that in period t, $L_i(t)$ and $L_{ij}(t)$ are determined contemporaneously in response to the prevailing coefficients $\alpha_{ij}(t)$. Agents then modify their coefficients in light of past performance. We may write

$$\alpha_{ij}(t+1) = g_i(\alpha_{ij}(t), L_j(t) - L_{ij}(t)), \tag{2.3}$$

where g_i is assumed to be nondecreasing in both arguments, bounded above by 1 and $g_i(a, 0) = a$.[11]

As mentioned in the Introduction, we consider three examples of *standards of appropriate behavior*: (1) agents' labor contributions are evaluated relative to the societal mean, (2) each agent considers what it would do in another agent's shoes, i.e., if it were in that agent's circumstances, and (3) agents are expected to contribute their fair share of labor.[12] We now consider each of these in turn.

3. Mean standard of behavior

In this section, we assume agents take the mean labor supply as the standard of behavior relative to which they judge the effort of others. That is, those who supply more labor than the mean are taken to be industrious and are held in greater esteem, while those who supply less than the mean are seen as shirking and garner less respect. Also, in this section we will, at times, restrict our attention to weakly benevolent agents, that is, those for whom α_{ij} is bounded below by 0. We will indicate when that is the case.

To demonstrate, we consider an example in which agents have identical direct

[11] Throughout most of the paper, we allow $\alpha_{ij} < 0$. However, in sections 3.2 and 3.3 we restrict our attention to weakly benevolent agents, that is, we impose the additional constraint $\alpha_{ij} \geq 0$.

[12] Here we assume that all agents employ the same standard of behavior, and we demonstrate several such standards. It would also be interesting to investigate the interactions among agents with different standards. That will be the subject of our future work.

preferences represented by $u(c, L) = c(1 - L)$. Substituting into (2.2) yields

$$U_i(\mathbf{L}) = \sum_j \alpha_{ij}\theta_j \prod_{k=1}^{n} L_k^{\beta_k}(1 - L_j). \tag{3.1}$$

Maximizing (3.1) with respect to L_i and simplifying, we obtain the following reaction function:[13]

$$L_i(L_{-i}) = \frac{\beta_i}{1 + \beta_i}\left(1 + \sum_{j \neq i} \alpha_{ij}\frac{\theta_j}{\theta_i}(1 - L_j)\right). \tag{3.2}$$

Note that under this specification of egoistic preferences, if agents were purely selfish, i.e., if there were no externality ($\alpha_{ij} = 0$, for all $j \neq i$), then i's labor supply would be $\frac{\beta_i}{1+\beta_i}$, which is constant (even taking into consideration the strategic effect of L_j on c_i) and, in particular, is independent of the output shares. Generally, however, according to (3.2) agent i's labor supply would consist of its egoistic utility maximizing level plus an additional term that reflects the spillover benefit on other agents.

Turning to the determination of $\boldsymbol{\alpha}$, as described above, the focus of our study is on fixed points of the (vector-valued) adjustment rules. Generally, let $x(t)$ denote the value of a variable x at time t, and let $\overline{x}(t)$ denote the average value of $x(t)$ among all agents.

Here, we take $L_{ij}(t) = \overline{L}(t)$; that is, at each point in time, each agent expects every other agent to supply the mean quantity of labor. Hence, we write

$$\alpha_{ij}(t + 1) = g(\alpha_{ij}(t), L_j(t) - \overline{L}(t)), \tag{3.3}$$

where g is subject to the aforementioned restrictions.

Substituting (3.3) into (3.2), we obtain

$$L_i(t + 1) = \frac{\beta_i}{1 + \beta_i}\left(1 + \sum_{j \neq i} g(\alpha_{ij}(t), L_j(t) - \overline{L}(t))\frac{\theta_j}{\theta_i}(1 - L_j(t + 1))\right). \tag{3.4}$$

We wish to focus on stationary Nash equilibria as determined by (3.4) and the corresponding stationary coefficients.[14] One type of stationary solution is that in

[13] Note that the second order condition $\frac{\partial^2 U_i}{\partial L_i^2} < 0$ is satisfied.

[14] In this paper we focus on the characterization of steady state outcomes, and we set aside the issue of stability of a steady state.

which all agents supply the mean quantity of labor. Alternatively, if we restrict our attention to (weakly) benevolent agents ($\alpha_{ij} \geq 0$, for all i, j), then there are two other types of stationary solutions as well. First, society might be partitioned into two clusters, A and B, where members of A supply labor above the mean and members of B supply below the mean. And second, in addition to A and B, there might be a third group, M, consisting of those who supply precisely the mean level of effort. In either partitioned outcome, it must be the case that for all i, $\alpha_{ij} = 1$ for all $j \in A$, and $\alpha_{ij} = 0$ for all $j \in B$, $j \neq i$.[15] We now consider each type in turn.

3.1. Equal effort steady states

We first consider solutions in which all agents supply the same quantity of labor. That is, let $(\mathbf{L}^*, \boldsymbol{\alpha}^*)$ be such a stationary Nash equilibrium together with the corresponding sentiments. Then $L_i^* = \overline{L}^*$ for all $i \in N$. From (3.4), we have

$$\overline{L}^* = \frac{\beta_i}{1+\beta_i}\left(1 + \sum_{j \neq i} \alpha_{ij}^* \frac{\theta_j}{\theta_i}(1 - \overline{L}^*)\right). \tag{3.5}$$

Let $\Lambda_{i\theta}^* \equiv \sum_j \alpha_{ij}^* \theta_j$ denote agent i's share-weighted concern for others, and let $T_{i\theta} \equiv \frac{\theta_i}{\beta_i}$. The expression $\Lambda_{i\theta}^*$ can be interpreted as the evaluation by agent i of the allocation of one unit of output. (We discuss the significance of $T_{i\theta}$ below.) From (3.5) we obtain

$$\frac{\overline{L}^*}{1 - \overline{L}^*} = T_{i\theta}^{-1} \Lambda_{i\theta}^* \tag{3.6}$$

and

$$\overline{L}^* = \frac{T_{i\theta}^{-1} \Lambda_{i\theta}^*}{1 + T_{i\theta}^{-1} \Lambda_{i\theta}^*}. \tag{3.7}$$

Notice that (3.6) is incompatible with $\Lambda_{i\theta}^* < 0$. Hence, even though we impose no lower bound on α_{ij} at this stage, there is a natural restriction on the aggregate extent of malevolence by each individual (at an interior steady state).

[15] We omit the dependence on t since these represent stationary values. Also, note that "clustering" in A and B is with respect to sentiments, not labor supply. That is, two agents in A might supply different quantities of labor (both greater than the mean), but each would earn maximal respect.

Since (3.6) holds for all $i \in N$, we have the following:

$$\frac{T_{i\theta}}{T_{j\theta}} = \frac{\Lambda^*_{i\theta}}{\Lambda^*_{j\theta}}. \tag{3.8}$$

Interpreting (3.8)[16], β_i might be considered the share of total output contributed by agent i. Therefore, if agents were rewarded on the basis of their marginal productivities, it would be the share received by i as well. Hence, $T_{i\theta}$ is a measure of the distance between the actual consumption allocation rule $\boldsymbol{\theta}$ and the marginal productivity (MP) or competitive rule $\boldsymbol{\beta}$. We refer to $T_{i\theta}$ as *i's treatment under $\boldsymbol{\theta}$*. Taking the MP reward as the benchmark, we say that agent *i is treated better than j under the allocation rule $\boldsymbol{\theta}$*, or that *$\boldsymbol{\theta}$ favors i over j*, if $T_{i\theta} > T_{j\theta}$. (Conversely, $T^{-1}_{i\theta}$ represents the amount i contributes relative to its reward and is referred to as i's *real contribution under $\boldsymbol{\theta}$*.) We have thus established the following:

Proposition 3.1. *In a steady state with equal effort, agents favored by the allocation rule must exhibit a greater degree of (share-weighted) concern for others.*

Indeed, according to (3.7), for each $\overline{L}^*$ there is a (convex) tradeoff between $\Lambda^*_{i\theta}$ and $T^{-1}_{i\theta}$ identifying the pairs that are consistent with $\overline{L}^*$ as a social norm. This is depicted in Figure 1. Note that as the common effort level increases, so, too, do the levels of concern and real contribution needed to support it ($\overline{L}'' > \overline{L}'$). Also, subject to the constraint $\Lambda^*_{i\theta} \leq 1$ (from $\alpha_{ij} \leq 1$) and the resource constraint on L_i, any of a continuum of social conventions might emerge.

Figure 1.

One implication of (3.8) is that if treatments are sufficiently asymmetric, then to to support steady state with equal effort some agents must be malevolent. Or conversely, the treatments must be sufficiently similar in order to support an equal effort steady state in which all agents are benevolent. This is easily seen for the case of two agents. It is then useful to compare the two-agent case with that of n agents.

First, let $n = 2$ and suppose $\theta_1 = 0.9$ ($\theta_2 = 0.1$) and $\beta_1 = 0.1$ ($\beta_2 = 0.9$). Then since $\alpha_{12} \leq 1$, it follows immediately from (3.8) that $\alpha_{21} < 0$. Indeed, in the

[16]Note that (3.6) and (3.8) impose no restriction on how each agent distributes its average concern over the other individuals, but depend only on the total. However, this is an artifact of the present functional forms and fails to be robust.

case of two agents, it is easy to identify all compatible treatments. Notice that (3.8) defines the following linear relationship between α_{12} and α_{21}:

$$\alpha_{21} = \frac{T_{2\theta}}{T_{1\theta}}\frac{\theta_2}{\theta_1}\alpha_{12} + \frac{\theta_2}{\theta_1}\left(\frac{\beta_1}{\beta_2} - 1\right). \tag{3.9}$$

Thus, we can characterize the parameter values for which α_{12} and α_{21} might be positive or negative and simultaneously for which $\alpha_{12}, \alpha_{21} \leq 1$. For example, taking $\boldsymbol{\theta}$ to be fixed, we can parameterize the consistent values in β_1. This is depicted in Figure 2.

Figure 2.

When $\beta_1 = 0.5$, the graph of (3.9) would pass through the origin and have slope $\left(\frac{\theta_2}{\theta_1}\right)^2$, as depicted by ℓ_1. As β_1 increases, the slope and intercept of (3.9) increase as well. For sufficiently large $\frac{\beta_1}{\beta_2}$, (3.9) is no longer consistent with $\alpha_{12} \geq 0$ when $\alpha_{21} \leq 1$. (ℓ_2 indicates the critical value.) Similarly, for sufficiently small $\frac{\beta_1}{\beta_2}$, $\alpha_{21} < 0$ when $\alpha_{12} \leq 1$. (ℓ_3 corresponds to the opposite critical value.) We state this formally as follows.

Proposition 3.2. *Let $n = 2$, and let $(\mathbf{L}^*, \boldsymbol{\alpha}^*)$ be a stationary outcome with equal effort. If $\frac{\beta_i}{\beta_j} > \frac{\theta_i}{\theta_j} + 1$, then $\alpha^*_{ij} < 0$. Conversely, if $\alpha^*_{12} \geq 0$ and $\alpha^*_{12} \geq 0$, then $\theta_1 \leq \frac{\beta_1}{\beta_2}$ and $\theta_2 \leq \frac{\beta_2}{\beta_1}$.*

Notice that the above proposition identifies sufficient conditions for one of the agents to be malevolent in a stationary outcome. While it is not possible for both $\frac{\beta_i}{\beta_j} > \frac{\theta_i}{\theta_j} + 1$ and $\frac{\beta_j}{\beta_i} > \frac{\theta_j}{\theta_i} + 1$ to hold simultaneously for $i \neq j$, nevertheless, as is obvious from Figure 2, it is indeed possible for both agents to be malevolent.

For the general case of $n > 2$ agents, the analogue of (3.9) is

$$\alpha_{ji} = \frac{T_{j\theta}}{T_{i\theta}}\frac{\theta_j}{\theta_i}\alpha_{ij} + \left[\frac{\theta_j}{\theta_i}\left(\frac{\beta_i}{\beta_j} - 1\right) + \frac{1}{\theta_i}\left(\sum\nolimits_{k\neq i,j}(\alpha_{ik} - \alpha_{jk})\theta_k\right)\right]. \tag{3.10}$$

Hence, we see that one agent's tendency to be malevolent toward another agent when treated relatively unfairly is modified by the latter's relative share-weighted concern for others. That is, for given $\frac{\beta_i}{\beta_j}$, agent j is less likely to be malevolent toward i if i is significantly more altruistic than j toward third parties.

Next, we examine the utility levels of the agents at a stationary outcome. By (3.1),

$$U_i(\mathbf{L}^*) = \overline{L}^*(1-\overline{L}^*)\Lambda^*_{i\theta}. \tag{3.11}$$

Using (3.6), we obtain

$$U_i(\mathbf{L}^*) = \frac{\theta_i}{\beta_i}\overline{L}^{*2}. \tag{3.12}$$

Hence, we obtain the striking result that under the *MP allocation rule* ($\theta_i = \beta_i$) the outcome is fully egalitarian in terms of social utilities.

The direct utility of individual i is

$$\theta_i Y^*(1-\overline{L}^*) = \theta_i \overline{L}^*(1-\overline{L}^*). \tag{3.13}$$

Therefore, the *egalitarian rule* ($\theta_i = \frac{1}{n}$) will equate egoistic utilities, irrespective of the productivities of the agents at the common effort level. Or, conversely, under the egalitarian allocation rule, social preferences will evolve in such a way so as to support the choice of a common effort level, irrespective of productivity differences. To sustain this rule, it must be the case that individuals with low productivities have more concern for others, as seen by (3.6).

3.2. 2-cluster stationary solutions

If we restrict our attention to weakly benevolent agents, then there are two other possible stationary solutions, namely, the clustered outcomes mentioned earlier. In this subsection we consider the case in which society partitions into two clusters consisting of those who work above average and garner maximal respect/concern and those who work below and garner the minimum. Again, let $(\mathbf{L}^*, \boldsymbol{\alpha}^*)$ be an interior stationary Nash equilibrium. Then by (3.2) we have the following:

$$L^*_a = \frac{\beta_a}{1+\beta_a}\left(1 + \frac{1}{\theta_a}\sum_{\substack{j\in A\\ j\neq a}} \theta_j(1-L^*_j)\right), \text{ for } a \in A, \tag{3.14}$$

$$L^*_b = \frac{\beta_b}{1+\beta_b}\left(1 + \frac{1}{\theta_b}\sum_{j\in A} \theta_j(1-L^*_j)\right), \text{ for } b \in B. \tag{3.15}$$

Define L_θ^{*A}, θ^A and β^A, respectively, by

$$\begin{aligned} L_\theta^{*A} &= \sum_{j\in A} \theta_j L_j^*, \\ \theta^A &= \sum_{j\in A} \theta_j, \\ \beta^A &= \sum_{j\in A} \beta_j. \end{aligned}$$

From (3.14) we obtain

$$L_a^* = \frac{\beta_a}{\theta_a}(\theta^A - L_\theta^{*A}). \tag{3.16}$$

Adding over all $a \in A$ and rearranging,

$$L_\theta^{*A} = \frac{\beta^A}{1+\beta^A}\theta^A. \tag{3.17}$$

Finally, using (3.15), (3.16) and (3.17), we obtain

$$L_a^* = \frac{\beta_a}{\theta_a}\frac{\theta^A}{1+\beta^A}, \tag{3.18}$$

$$L_b^* = \frac{\beta_b}{1+\beta_b}\left(1 + \frac{1}{\theta_b}\frac{\theta^A}{1+\beta^A}\right). \tag{3.19}$$

From the labor supply expressions (3.18) and (3.19), we see that L_i^* is generally increasing in β_i and decreasing in θ_i, for $i = a, b$.[17] The former is due to the fact that an increase in β_i increases the return to L_i, both in its effect on i's own consumption as well as the spillover effect on other agents' consumption. The latter, on the other hand, is purely a spillover effect. As noted earlier, if the externality were not present, (i.e., if $U_i(\mathbf{c}, \mathbf{L}) = u_i(c, L)$), then labor supply would be independent of θ_i. Hence, the fact that, here, labor supply does vary with θ_i is due solely to the fact that an increase in θ_i decreases the share of output received by the other agents and thus decreases the (external) return to one's labor effort.

Also from (3.16) or (3.18), we see that for given β^A and θ^A, the labor supplies among those in A are arrayed linearly in either the agents' real contributions $T_{i\theta}^{-1}$

[17] By assumption, $L_b^* < \overline{L}^* < L_a^*$, for all $a \in A$ and $b \in B$.

or, if we take $\boldsymbol{\theta}$ to be fixed, then in their productivities.[18] Also, within A, we have

$$\frac{L_a^*}{L_{a'}^*} = \frac{T_{a\theta}^{-1}}{T_{a'\theta}^{-1}}, \text{ for } a, a' \in A. \tag{3.20}$$

That is, labor supply varies inversely with treatment: those who are treated well under the allocation rule $\boldsymbol{\theta}$ (i.e., $\frac{\theta_a}{\beta_a}$ is large) supply less labor than those who are treated poorly.

To further explore the labor supply behavior in (3.18) and (3.19), suppose $\boldsymbol{\theta}$ were fixed. Then, as mentioned above, for given β^A and θ^A, L_a^* is linear in β_a, while L_b^* is strictly concave in β_b. Thus, as depicted in Figure 3, there may be three possible configurations between the labor supply curves (3.18) and (3.19).

Figure 3.

It may be the case that L_b^* is greater than L_a^* for all β, as in ℓ_1; that L_b^* is less than L_a^* for all β, as in ℓ_2; or that L_b^* is greater than L_a^* for small β and less for large β, as in ℓ_3. Conditions for each of these are easily determined by comparing the slopes of the respective labor supply schedules at zero and/or one.

Usually, in a 2-cluster steady state, the members of A are those with high productivity. Indeed that must be the case for the labor supply curve ℓ_1. That is, comparing ℓ_1 to L_a^*, it is impossible for agents with greater productivity supplying labor according to L_a^* to work less than those with lower productivity supplying labor according to ℓ_1. However, the other cases (ℓ_2 and ℓ_3) present the interesting possibility that the low productivity agents work more than those with high productivity in a steady state.[19] For example, suppose society were partitioned such that agents had either low productivity, β_l, or high productivity, β_h, and the former supplied labor at point c in Figure 3 while the latter supplied either at d if the labor supply curve were ℓ_2 or e if it were ℓ_3.

Proposition 3.3. *In a two-cluster steady state, the productivity ranking of the agents is ambiguous. That is, it is possible that the high productivity agents work more than the low productivity agents or vice versa.*

[18] For $a \in A$, the individual labor supply is not a linear function of $T_{a\theta}^{-1}$ or β_a since, generally, changes in β_a or θ_a affect either β^A, θ^A or both. However, for the coalition A as a whole, β^A and θ^A are fixed by definition. Therefore, for distinct agents $a, a' \in A$, their respective labor supplies do indeed lie along (3.18).

[19] Consider minimum wage workers who hold two or more jobs in order to make ends meet. According to our model, they might garner more respect than those who work less but earn more.

PROOF. The proof is by example. We omit demonstrating that A might contain those with high productivity since that is the norm. To show that the reverse is possible, let $N = \{1, 2, 3\}$[20], and suppose $A = \{1, 2\}$ and $B = \{3\}$. Agents 1 and 2 both have productivity $\beta_a = 0.3$ and each receives an output share of $\theta_a = 0.1$. Hence, $\theta^A = 0.2$ and $\beta^A = 0.6$, and therefore $\beta_3 = 0.4$ and $\theta_3 = 0.8$. From (3.18) and (3.19), we have $L_1^* = L_2^* = 0.375$ and $L_3^* = 0.33$, and thus $\overline{L} = 0.36$. (This corresponds to a configuration such as c and e in Figure 3, above.) ■

Turning to the welfare of the two groups, it is easy to show that all members of A enjoy the same extended utility,

$$U_a(\mathbf{L}^*) = \frac{\theta^A}{1+\beta^A} Y^*.$$

As for $b \in B$, we have

$$U_b(\mathbf{L}^*) = \theta_b Y^* + U_a(\mathbf{L}^*) = \left(\theta_b + \frac{\theta^A}{1+\beta^A}\right) Y^* > U_a(\mathbf{L}^*).$$

Hence, those individuals who work less than the mean (and hence garner less respect/esteem) will obtain higher extended utility than those held in greater esteem.

3.3. 3-cluster solutions

Finally, we briefly consider three cluster equilibria. Again, let $(\mathbf{L}^*, \boldsymbol{\alpha}^*)$ be an interior stationary Nash equilibrium. From (3.2) we have

$$L_a^* = \frac{\beta_a}{1+\beta_a}\left(1 + \frac{1}{\theta_a}\sum_{j\in M} \alpha_{aj}^* \theta_j (1-\overline{L}^*) + \frac{1}{\theta_a}\sum_{\substack{j\in A \\ j\neq a}} \theta_j (1 - L_j^*)\right), \text{ for } a \in A, \tag{3.21}$$

$$L_m^* = \overline{L}^* = \frac{\beta_m}{1+\beta_m}\left(1 + \frac{1}{\theta_m}\sum_{\substack{j\in M \\ j\neq m}} \alpha_{mj}^* \theta_j (1-\overline{L}^*) + \frac{1}{\theta_m}\sum_{j\in A} \theta_j (1 - L_j^*)\right), \text{ for } m \in M, \tag{3.22}$$

[20] One can show that this requires more than two agents.

$$L_b^* = \frac{\beta_b}{1+\beta_b}\left(1 + \frac{1}{\theta_b}\sum_{j\in M}\alpha_{bj}^*\theta_j(1-\overline{L}^*) + \frac{1}{\theta_b}\sum_{j\in A}\theta_j(1-L_j^*)\right), \text{ for } b \in B. \quad (3.23)$$

In keeping with the earlier notation, for $S \subseteq N$, let $\Lambda_{i\theta}^{*S} = \sum_{j\in S}\alpha_{ij}^*\theta_j$ denote i's aggregate share-weighted concern for the members of S. In the event $i \in S$, let $\Lambda_{i\theta}^{*S_{-i}} = \sum_{\substack{j\in S\\ j\neq i}}\alpha_{ij}^*\theta_j$ denote i's weighted concern for the other members. Then from (3.21) we obtain

$$L_a^* = \frac{\beta_a}{\theta_a}(\theta^A - L_\theta^{*A} + (1-\overline{L}^*)\Lambda_{a\theta}^{*M}). \quad (3.24)$$

In contrast to (3.16), now the agent's concern for the members of the middle group enters positively in its labor supply decision. Consequently, within A, the agents' *relative* labor contributions will depend on their relative concern as well, rather than solely on their relative treatments (see (3.20)).

Aggregating (3.24) over A, we then establish

$$L_\theta^{*A} = \frac{1}{1+\beta^A}(\beta^A\theta^A + (1-\overline{L}^*)\sum_{j\in A}\beta_j\Lambda_{j\theta}^{*M}). \quad (3.25)$$

Substituting this into (3.24) and simplifying yields

$$L_a^* = \frac{\beta_a}{\theta_a}\frac{1}{1+\beta^A}\left(\theta^A + (1-\overline{L}^*)\Lambda_{a\theta}^{*M} + (1-\overline{L}^*)\sum_{j\in A}\beta_j(\Lambda_{a\theta}^{*M} - \Lambda_{j\theta}^{*M})\right). \quad (3.26)$$

Thus, the labor supply decision by $a \in A$ takes into consideration three factors: the impact on the members of A (as in (3.17)), the impact on the members of M weighted by the extent of a's concern for that group, and a third term which reflects the comparative concern by a for M versus that of the other members of A. The greater the concern for M by others, the less a need contribute to the well-being of its members. The latter term thus captures the free-rider effect in generating spillover benefits *for* M.

Analogously, we obtain the following expressions for L_b^* and L_m^*, respectively:

$$L_b^* = \frac{\beta_b}{1+\beta_b}\left(1 + \frac{1}{\theta_b}\frac{\theta^A}{1+\beta^A} + \frac{1}{\theta_b}(1-\overline{L}^*)\Lambda_{b\theta}^{*M} - \frac{1}{\theta_b}\frac{1}{1+\beta^A}(1-\overline{L}^*)\sum_{j\in A}\beta_j\Lambda_{a\theta}^{*M}\right) \quad (3.27)$$

$$L_m^* = \overline{L}^* = \frac{\beta_m}{1+\beta_m}\left(1 + \frac{1}{\theta_m}\frac{\theta^A}{1+\beta^A} + \frac{1}{\theta_m}(1-\overline{L}^*)\Lambda_{m\theta}^{*M-m} - \frac{1}{\theta_m}\frac{1}{1+\beta^A}(1-\overline{L}^*)\sum_{j\in A}\beta_j\Lambda_{a\theta}^{*M}\right). \tag{3.28}$$

In contrast to (3.26), (3.27) and (3.28) both contain an additional term reflecting the agent's concern for its own well-being. Also, the final term in each expression no longer depends on the agent's comparative concern for M, and thus represents an absolute (versus relative) free-rider effect.

4. Transposition, or "If I were you..."

In this section, we consider an alternative standard of behavior in which agents determine what they think is appropriate behavior for others by asking what they themselves would do in similar circumstances. This seems particularly appealing in describing how people with diverse characteristics might evaluate the behavior of those with whom they have little in common. For instance, it is easy to imagine asking oneself how you would behave if you were a member of a disenfranchised minority, or if you were disabled, or if you had amassed or inherited great wealth, etc., if none of these were actually the case.

To begin, we describe a procedure for transposing circumstances.[21] Then we focus on particular examples in order to explore possible implications of the alternative rule.

First, recall that to determine L_i, agent i solves the program

$$P_i: \quad \max_{L_i} U_i(\mathbf{L}). \tag{4.1}$$

The vector of actual behaviors, $\mathbf{L}$, is determined through the simultaneous solution of (4.1) for all $i \in N$. Let $\widetilde{\mathbf{L}} \equiv (\widetilde{L}_i)_{i\in N}$ denote the resulting labor supply vector.

Next, we consider agent i's evaluation of the appropriateness of $\widetilde{L}_j$. To do so, we ask how would i solve the problem confronting agent j? One way to formulate this counterfactual is as follows.

Expanding U_j,

$$U_j(\mathbf{L}) = u_j(\theta_j f(\mathbf{L}), L_j) + \alpha_{ji}u_i(\theta_i f(\mathbf{L}), L_i) + \sum_{k\neq i,j}\alpha_{jk}u_k(\theta_k f(\mathbf{L}), L_k). \tag{4.2}$$

[21] We would point out that there may be other plausible procedures.

From i's perspective, i.e., evaluated according to its own (direct and social) preferences, (4.2) might appear as

$$U_{ij}(\mathbf{L}) = u_i(\theta_j f(\mathbf{L}), L_j) + \alpha_{ij} u_i(\theta_i f(\mathbf{L}), L_i) + \sum_{k \neq i,j} \alpha_{ik} u_k(\theta_k f(\mathbf{L}), L_k). \tag{4.3}$$

In other words, if i were in j's position, it might use its own direct preferences to evaluate the impact of its decision on itself, and its own coefficients α_i to evaluate the impact on others. Also, in constructing U_{ij} we affix the coefficient α_{ij} to $u_i(\theta_i f(\mathbf{L}), L_i)$. That is, for the purpose of the example, we assume that i considers what would be appropriate behavior for j if j were to care about i to the same extent that i cares about j?[22]

For L_{ij}, i would solve

$$P_{ij}: \quad \max_{L_j} U_{ij}(\mathbf{L}). \tag{4.4}$$

Solving the $n-1$ programs P_k, for $k \neq j$, together with P_{ij} yields Nash equilibrium (i.e., consistent) behaviors, which we denote $\widehat{\mathbf{L}}^i$. We take $L_{ij} = \widehat{L}^i_j$; that is, $\widehat{L}^i_j$ is the optimal quantity of labor i would supply if it were j, as described above, and if other agents were to behave optimally according to their actual extended utilities.

As before, the discrepancy between j's actual Nash equilibrium behavior, $\widetilde{L}_j$ and i's evaluation of appropriate behavior for j, namely, $\widehat{L}^i_j$, forms the basis for its revising α_{ij}. The interaction is then repeated relative to the new altruism coefficients to yield revised (actual) behaviors and standards for others. Etc. We again denote steady state labor supplies by $\mathbf{L}^* \equiv (L^*_i)_{i \in N}$.

To demonstrate the consequences of this standard of behavior, we again consider a two agent example. However, since $\widetilde{L}_j$ and $\widehat{L}^i_j$ might differ for various reasons – including differences in tastes, abilities (and in other contexts resources) – we consider such characteristics separately in order to explore the implications of each. That is, we consider two examples, one in which agents have identical preferences but different abilities, and a second in which their abilities are the same, but their preferences differ.

[22] Here, we are considering the concern i would have for the (hypothetical) person occupying its (own) original position, were it to occupy the position of j. Such a thought experiment is purely hypothetical and there need be no correct assignation. In any event, our method will suffice to demonstrate some of the possible implications of the theory.

4.1. Productivity differences

Returning to the example of the previous section, we consider the simple case involving two agents who have identical private preferences represented by $u(c, L) = c(1 - L)$.

First, solving (4.1), we obtain the (actual) labor supply decisions as in (3.2):

$$L_i = \frac{\beta_i}{1+\beta_i}\left(1 + \alpha_{ij}\frac{\theta_j}{\theta_i}(1 - L_j)\right), \text{ for } i, j = 1, 2, i \neq j. \tag{4.5}$$

This leads to the Nash equilibrium behaviors:

$$\widetilde{L}_i = \frac{\beta_i}{\theta_i}\left(\frac{(1+\beta_j)\theta_i - \alpha_{ij}\alpha_{ji}\beta_j\theta_i + \alpha_{ij}\theta_j}{(1+\beta_i)(1+\beta_j) - \alpha_{ij}\alpha_{ji}\beta_i\beta_j}\right). \tag{4.6}$$

Alternatively, from (4.4) and (4.1) we obtain

$$L_{ij} = \frac{\beta_j}{1+\beta_j}\left(1 + \alpha_{ij}\frac{\theta_i}{\theta_j}(1 - L_i)\right). \tag{4.7}$$

Using (4.5) and (4.7), we obtain the following standards[23]:

$$\widehat{L}^i_j = \frac{\beta_j}{\theta_j}\left(\frac{(1+\beta_i)\theta_j - (\alpha_{ij})^2\beta_i\theta_j + \alpha_{ij}\theta_i}{(1+\beta_i)(1+\beta_j) - (\alpha_{ij})^2\beta_i\beta_j}\right). \tag{4.8}$$

Again assuming the agents adjust their α coefficients in accordance with (2.3), an interior steady state associated with (4.6) and (4.8) requires, in general, that $\alpha_{12} = \alpha_{21} = \alpha^*$ and entails the labor contributions

$$\begin{aligned} L^*_1 &= \frac{\beta_1}{\theta_1}\left(\frac{(1+\beta_2)\theta_1 - \alpha^{*2}\beta_2\theta_1 + \alpha^*\theta_2}{(1+\beta_1)(1+\beta_2) - \alpha^{*2}\beta_1\beta_2}\right) \\ L^*_2 &= \frac{\beta_2}{\theta_2}\left(\frac{(1+\beta_1)\theta_2 - \alpha^{*2}\beta_1\theta_2 + \alpha^*\theta_1}{(1+\beta_1)(1+\beta_2) - \alpha^{*2}\beta_1\beta_2}\right). \end{aligned} \tag{4.9}$$

Thus, a steady state involves equal (reciprocal) degrees of concern but possibly different labor contributions based on the agents' exogenous characteristics or circumstances. Also, a range of (mutual) stationary sentiments might occur. For

[23]Since, here, agents have the same egoistic preferences, $\widetilde{L}_j$ and $\widehat{L}^i_j$ differ only in the welfare weights ascribed to the other agent.

example, when $\alpha^* = 0$, $L_i^* = \frac{\beta_i}{1+\beta_i}$, for $i = 1, 2$, as pointed out earlier. Also, for $\alpha^* = 1$, $L_i^* = \frac{\beta_i}{2\theta_i}$. In this case, if $\theta_1 = \theta_2 = \frac{1}{2}$, then $L_i^* = \beta_i$. This seems to confirm our intuition that if agents exhibit a greater degree of mutual concern, they contribute more to the common good ($\beta_i > \frac{\beta_i}{1+\beta_i}$). However, as we shall see that need not be the case in general.

Exploring the general behavior of L_i^* in (4.9) as a function of θ_i, β_i and α^*, we first note that L_i^* is decreasing in θ_i.[24] This is again a pure spillover effect, as discussed earlier.

Turning to the role of β_i, first, the direct effect of an increase in β_i in (4.5) is to increase the return to labor, both in its impact on one's own private utility as well as the spillover effect on the other agent's private utility. Moreover, an increase in β_i will be accompanied by a decrease in β_j, which, in a similar manner, will decrease the (direct) return to L_j. Since L_j is negatively related to L_i^*, the indirect effect on L_i is positive as well. Hence, the overall effect of an increase in β_i is unambiguously positive.[25]

Finally, and most interestingly, we consider the role of α^*. As with β_i, the direct effect of α^* on L_i in (4.5) is positive. Now, however, α^* has a positive effect on L_j as well, which as mentioned above, negatively affects L_i^*. Hence, the overall effect on L_i^* is ambiguous. Generally, whether the direct or the indirect effect dominates depends on the circumstances of the agents, i.e., on $(\boldsymbol{\beta}, \boldsymbol{\theta})$. To demonstrate, consider the case in which $\beta_1 = \beta_2 = 0.5$. From (4.9), we have

$$L_1^* = \frac{0.5}{\theta_1 \left(2.25 - 0.25\alpha^2\right)} \left(1.5\theta_1 + \alpha - \theta_1\alpha - 0.5\theta_1\alpha^2\right). \tag{4.10}$$

The graph of (4.10) is depicted in Figure 4.

[24] Note that L_i^* can be written as $a + b(\frac{1-\theta_i}{\theta_i})$, where a and b are independent of θ_i.

[25] Here, too, it can be shown analytically that $\frac{\partial L_i^*}{\partial \beta_i} > 0$ in the region $\alpha, \beta_i \in [0, 1]$. However, the details are quite tedious.

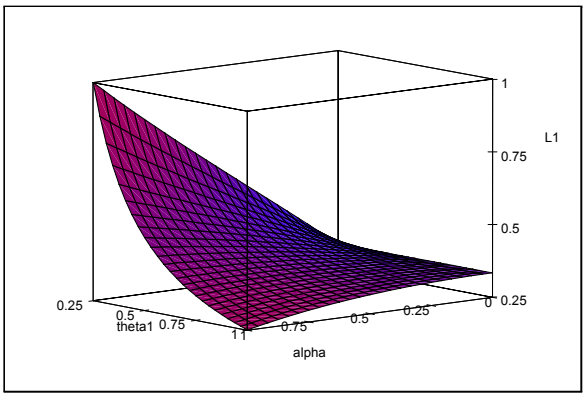


Figure 4. L_1^* as a function of θ_1 and α, for $\beta_1 = \beta_2 = 0.5$.

Notice that the effect of α on L_1^* differs significantly for large θ_1 versus small. (Indeed, it is unambiguously the case that $\frac{\partial^2 L_1^*}{\partial \theta_1 \partial \alpha} < 0$.) For instance, from (4.10), at $\theta_1 = 0.8$, we have

$$L_1^* = \frac{0.5}{1.8 - 0.2\alpha^2}\left(0.2\alpha - 0.4\alpha^2 + 1.2\right),$$

which has the following graph:

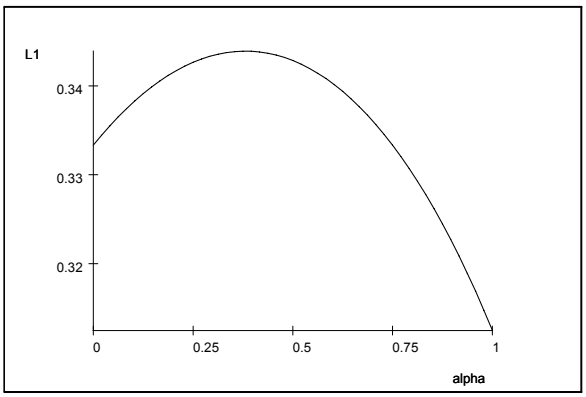


Figure 5. L_1^* as a function of α, for $\beta_1 = \beta_2 = 0.5$ and $\theta_1 = 0.8$.

Thus, it is possible that as 1's concern for 2 increases, 1 might actually reduce its labor supply!

Alternatively, at $\theta_1 = 0.4$,

$$L_1^* = \frac{0.5}{0.9 - 0.1\alpha^2}\left(0.6\alpha - 0.2\alpha^2 + 0.6\right),$$

which is uniformly increasing in α, for all $\alpha \in [0,1]$. Indeed, for each $\beta_1 \in [0,1]$, there is a critical value of θ_1 below which L_1^* is monotonic in α and above which it is nonmonotonic. This critical value is given by the expression

$$\theta_1^* = \frac{1}{3-2\beta_1}\left(1+\beta_1-\beta_1^2\right), \tag{4.11}$$

which is increasing in β_1 over $[0,1]$. That is, the greater is β_1, the broader the range of θ_1 over which L_1^* is monotonically increasing in α. In the extreme case of $\beta_1 = 1$, this critical value is 1; in this case L_1^* is increasing in α for all θ_1.

Intuitively, the effect of θ_1 on $\frac{\partial L_1^*}{\partial \alpha}$ is similar to the effect of θ_1 directly on L_1^*: the smaller is θ_1 (i.e., the larger is θ_2), the greater the marginal benefit derived from the spillover and hence the greater is L_1^*. Moreover, a decrease in θ_1 amplifies the impact of α on L_1^*.

We encounter a similar anomaly in the effect of α on L_1^* when we reverse the roles of θ_1 and β_1. For instance, suppose $\theta_1 = 0.9$. Then from (4.9), we have

$$L_1^* = \frac{\beta_1}{\beta_1-\beta_1^2-\alpha^2\beta_1+\alpha^2\beta_1^2+2}\left(0.11111\alpha-1.0\beta_1-1.0\alpha^2+1.0\alpha^2\beta_1+2.0\right). \tag{4.12}$$

This is depicted in Figure 6.

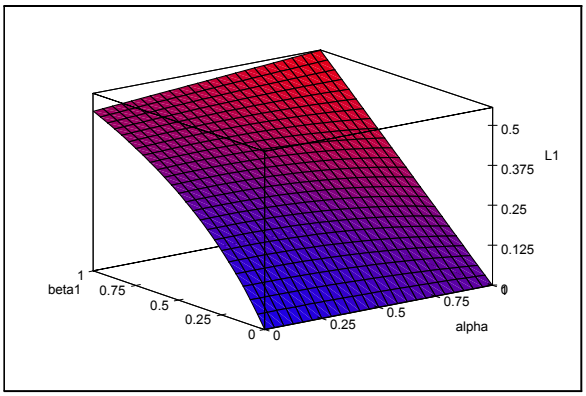


Figure 6. L_1^* as a function of β_1 and α, for $\theta_1 = 0.9$.

Here, too, the behavior is generally not uniform. For example, at $\beta_1 = 0.95$, (4.12) becomes

$$L_1^* = \frac{0.95}{2.0475-0.0475\alpha^2}\left(0.11111\alpha-0.05\alpha^2+1.05\right),$$

which exhibits the following monotonic behavior:

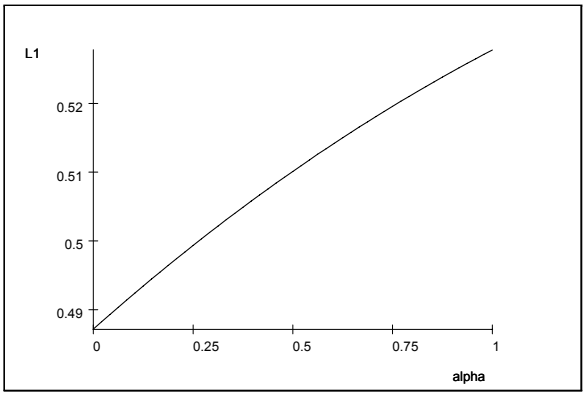


Figure 7. L_1^* as a function of α, for $\theta_1 = 0.9$ and $\beta_1 = 0.95$.

However, at $\beta_1 = 0.8$, we have

$$L_1^* = \frac{0.8}{2.16 - 0.16\alpha^2}\left(0.11111\alpha - 0.2\alpha^2 + 1.2\right),$$

which has the following graph:

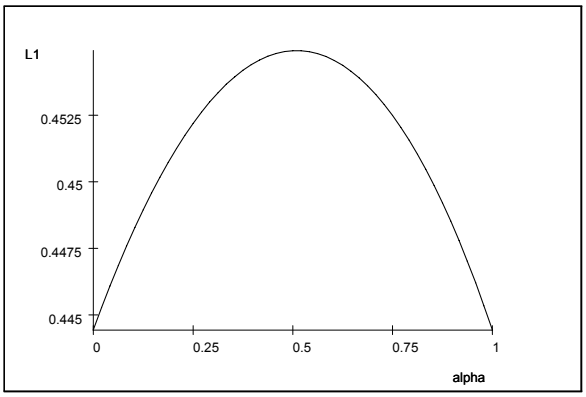


Figure 8. L_1^* as a function of α, for $\theta_1 = 0.9$ and $\beta_1 = 0.8$.

As with θ_1, we can identify the values of β_1 over which L_1^* is monotonic in α. This is done by inverting (4.11). Here, however, the regions are reversed relative to θ_1; L_1^* is nonmonotonic for low values of β_1 and monotonically increasing for high values. This is summarized in the following figure, which depicts the graph of the critical equation (4.11) for $\theta_1, \beta_1 \in [0, 1]$. For pairs (β_1, θ_1) lying below the graph, L_1^* is monotonic in α; and for those lying above, L_1^* is nonmonotonic.

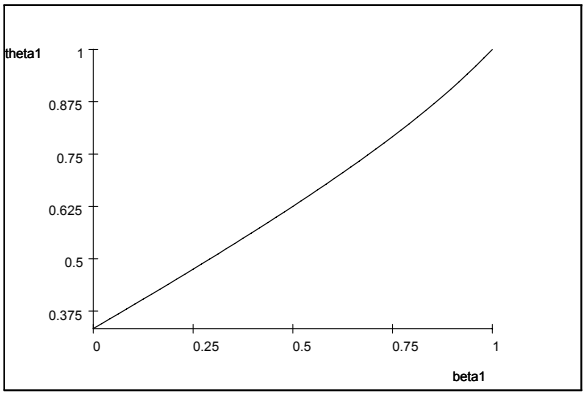


Figure 9.

Note that while L_1^* can decrease with α, it cannot be the case that an increase in α leads to a decrease in both L_1^* and L_2^* simultaneously. Indeed, it is easily shown that if $\theta_1 > \theta_1^*$ in (4.11), then $\theta_2 < \theta_2^*$ for the corresponding critical value of θ_2. Similarly, the regions of β_1 and β_2 over which the nonmonotonicity occurs do not overlap.

From (4.5) it is clear that in every case the direct effect of an increase in α is to increase labor supply. Thus, in general, such counterintuitive behavior arises because, in the relevant parameter range, the indirect effect (i.e., the effect on the other agent's labor supply and the effect of that on one's own) dominates the direct effect. This is most pronounced either when one agent appropriates the lion's share of output (θ_i is large and therefore the spillover benefit of L_i is small) or when the other agent is significantly more productive (β_j is large). Regarding the latter, suppose an increase in L_i would (indirectly) cause L_j to decrease. Then if β_j is large, this might significantly reduce overall output and hence the portion allocated to j – especially if θ_j is small. Indeed, the adverse effect on output of the decrease in L_j might more than offset the positive effect of the increase in L_i. As a result, j might end up worse off (in terms of its egoistic preferences). Therefore, the opposite behavior, decreasing L_i, might actually make j better off.

Next, for the purpose of comparison, we contrast the outcomes for the two agents under the MP allocation rule ($\theta_i = \beta_i$) versus the egalitarian rule ($\theta_i = \frac{1}{2}$),

for arbitrary α. First, under the MP rule, (4.9) can be written as

$$\begin{aligned} L_1^* &= \left(\frac{(\beta_2+1)\beta_1 - \alpha^2\beta_2\beta_1 + \alpha\beta_2}{(\beta_1+1)(\beta_2+1) - \alpha^2\beta_1\beta_2}\right) \\ L_2^* &= \left(\frac{(\beta_1+1)\beta_2 - \alpha^2\beta_1\beta_2 + \alpha\beta_1}{(\beta_1+1)(\beta_2+1) - \alpha^2\beta_1\beta_2}\right). \end{aligned}$$

In this case it is easy to show that the equilibrium labor contributions and egoistic utilities are ranked in the same order as the agents' productivities, that is, $L_1^* \geq L_2^*$ and $u_1(\theta_1 L_1^{*\beta_1} L_2^{*\beta_2}, L_1^*) \geq u_2(\theta_2 L_1^{*\beta_1} L_2^{*\beta_2}, L_2^*)$ iff $\beta_1 \geq \beta_2$. Moreover, since the agents exhibit the same degree of concern, their social utilities are ranked in the same fashion.

Alternatively, under the egalitarian rule, (4.9) simplifies as follows:

$$\begin{aligned} L_1^* &= \beta_1\left(\frac{(\beta_2+1) - \alpha^2\beta_2 + \alpha}{(\beta_1+1)(\beta_2+1) - \alpha^2\beta_1\beta_2}\right) \\ L_2^* &= \beta_2\left(\frac{(\beta_1+1) - \alpha^2\beta_1 + \alpha}{(\beta_1+1)(\beta_2+1) - \alpha^2\beta_1\beta_2}\right). \end{aligned}$$

Here, the low productivity agent works less and, since $\theta_1 = \theta_2$, derives greater egoistic and social utility.

4.2. Taste differences

We now turn to a second example in which agents have identical productivities but different tastes. Again, we consider the case of two agents (hence, $\beta_1 = \beta_2 = \frac{1}{2}$), but now we take their egoistic preferences to be represented by $u_1(c, L) = c^2(1-L)$ and $u_2(c, L) = c^2(1-2L)$, respectively. Thus, agent 2 has a taste bias for leisure relative to agent 1.

As above, the agents' actual behaviors are determined by solving:

$$\max_{L_1} \theta_1^2 L_1 L_2(1-L_1) + \alpha_{12}\theta_2^2 L_1 L_2(1-2L_2), \tag{4.13}$$

$$\max_{L_2} \theta_2^2 L_1 L_2(1-2L_2) + \alpha_{21}\theta_1^2 L_1 L_2(1-L_1). \tag{4.14}$$

This yields the reaction functions

$$L_1 = \frac{\theta_1^2 + \alpha_{12}\theta_2^2(1-2L_2)}{2\theta_1^2}, \tag{4.15}$$

$$L_2 = \frac{\theta_2^2 + \alpha_{21}\theta_1^2(1-L_1)}{4\theta_2^2}. \tag{4.16}$$

Solving (4.15) and (4.16) simultaneously yields the Nash equilibrium (actual) labor supplies

$$\widetilde{L}_1 = \frac{2\theta_1^2 + \alpha_{12}\theta_2^2 - \alpha_{12}\alpha_{21}\theta_1^2}{(4 - \alpha_{12}\alpha_{21})\theta_1^2}, \tag{4.17}$$

$$\widetilde{L}_2 = \frac{2\theta_2^2 + \alpha_{21}\theta_1^2 - \alpha_{12}\alpha_{21}\theta_2^2}{2(4 - \alpha_{12}\alpha_{21})\theta_2^2}. \tag{4.18}$$

Next, to obtain $\widehat{L}_2^1$ we solve (4.15) together with the following program which describes 1's perception of the decision problem facing agent 2:

$$\max_{L_2} \theta_2^2 L_1 L_2(1 - L_2) + \alpha_{12}\theta_1^2 L_1 L_2(1 - L_1). \tag{4.19}$$

This yields the reaction function

$$L_2^1 = \frac{\theta_2^2 + \alpha_{12}\theta_1^2(1 - L_1)}{2\theta_2^2}, \tag{4.20}$$

which together with (4.15) yields

$$\widehat{L}_2^1 = \frac{2\theta_2^2 + \alpha_{12}\theta_1^2 - \alpha_{12}^2\theta_2^2}{2\theta_2^2(2 - \alpha_{12}^2)}. \tag{4.21}$$

Similarly, from

$$\max_{L_1} \theta_1^2 L_1 L_2(1 - L_1) + \alpha_{12}\theta_{12}^2 L_1 L_2(1 - L_2)$$

we obtain the reaction function

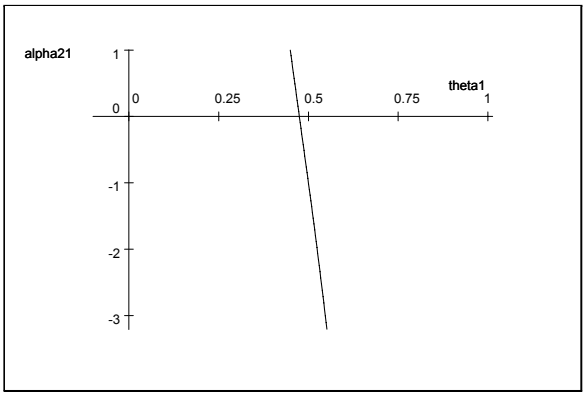
$$L_1^2 = \frac{\theta_1^2 + \alpha_{21}\theta_2^2(1 - 2L_2)}{4\theta_1^2}. \tag{4.22}$$

And this together with (4.16) yields

$$\widehat{L}_1^2 = \frac{2\theta_1^2 + \alpha_{21}\theta_2^2 - \alpha_{21}^2\theta_1^2}{\theta_1^2(8 - \alpha_{21}^2)}. \tag{4.23}$$

Equating $\widehat{L}_1^2$ to $\widetilde{L}_1$ and $\widehat{L}_2^1$ to $\widetilde{L}_2$, we obtain the stationary coefficients. These equations have four roots, only two of which are feasible given the restrictions that

$\alpha_{12} \leq 1$ and $\alpha_{21} \leq 1$. The feasible solutions are $\alpha_{12} = -\sqrt{2}$, $\alpha_{21} = -2\sqrt{2}$ and, for $\theta_1 \geq 0.449$, $\alpha_{12} = -\frac{\theta_1^2}{\theta_2^2}$, $\alpha_{21} = \left(\frac{2\theta_2^2}{\theta_1^2} - \frac{3\theta_1^2}{\theta_2^2}\right)$. It is interesting to note that agent 1 is always malevolent in a steady state, even when θ_1 is large, and agent 2 may be either benevolent or malevolent. To be precise, the (feasible) stationary values of α_{21} for each θ_1 are depicted in the following figure.[26] Thus, α_{21} is generally negative, although there is a small region (approximately for $\theta_1 \in (0.4494, 0.4747)$) in which it is positive.



Note that for the case of $\theta_1 = \theta_2 = \frac{1}{2}$, the stationary coefficients are $\alpha_{12} = \alpha_{21} = -1$.[27]

5. Fair share

Finally, we consider one additional standard in which agents believe that each person should contribute their "fair share" of labor. While there are various interpretations of "fair share," here we take it to mean that each person expects others to contribute labor in proportion to their output shares.[28] Thus, relative

[26] For $\theta_1 \in (0, 0.449)$, the stationarity conditions do yield consistent α's. However, α_{21} exceeds 1. (In fact, α_{21} is unbounded on the interval).

[27] Since $\beta_1 = \beta_2$, this corresponds to both the egalitarian and the MP rule.

[28] This does not preclude consideration of productivity differences since that may be taken into consideration in determining the division rule $\boldsymbol{\theta}$. Indeed, the example might be quite fitting in the context of a partnership where profit shares are negotiated and contractually predetermined but labor inputs are ongoing.

to agent i's own labor supply, L_i, L_{ij} is defined implicity by

$$\frac{L_i}{L_{ij}} = \frac{\theta_i}{\theta_j}. \tag{5.1}$$

Again, for the purpose of demonstration we consider the case of two agents with identical private preferences $u(c, 1-L) = c(1-L)$.

First, given α_{12} and α_{21}, the actual Nash equilibrium labor supplies $\widetilde{\mathbf{L}}$ are given by (4.6). Substituting into (5.1), we have that at a stationary outcome (i.e., $\widetilde{L}_j = L_{ij}$),

$$\frac{\beta_1\left((1+\beta_2) - (\alpha^*_{12})(\alpha^*_{21})\beta_2 + \alpha^*_{12}\frac{\theta_2}{\theta_1}\right)}{\beta_2\left((1+\beta_1) - (\alpha^*_{21})(\alpha^*_{12})\beta_1 + \alpha^*_{21}\frac{\theta_1}{\theta_2}\right)} = \frac{\theta_1}{\theta_2}.$$

Or,

$$\frac{\left((1+\beta_2) - (\alpha^*_{12})(\alpha^*_{21})\beta_2 + \alpha^*_{12}\frac{\theta_2}{\theta_1}\right)}{\left((1+\beta_1) - (\alpha^*_{21})(\alpha^*_{12})\beta_1 + \alpha^*_{21}\frac{\theta_1}{\theta_2}\right)} = \frac{\frac{\theta_1}{\beta_1}}{\frac{\theta_2}{\beta_2}}. \tag{5.2}$$

In general, the patterns of utility interdependence consistent with (5.2) are quite complex. Therefore, we focus only on special cases.

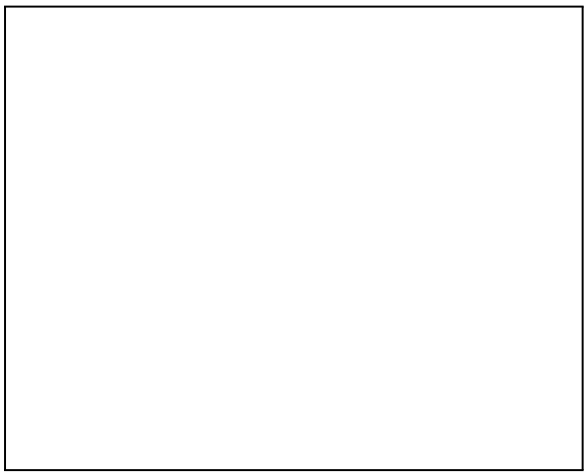
Proposition 5.1. *If $\theta_i = \beta_i = \frac{1}{2}$ for $i = 1, 2$, then any mutual pattern of concern, $\alpha^*_{12} = \alpha^*_{21} \in (-\infty, 1]$ is stable.*

More generally, under the MP allocation rule (5.2) specializes as follows:

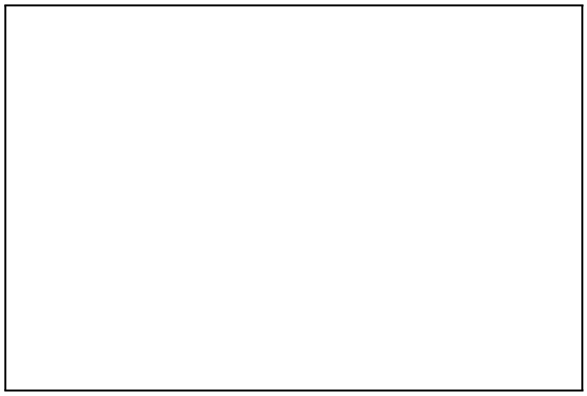
$$(\beta_2 - \beta_1)(1 - (\alpha^*_{12})(\alpha^*_{21})) = \alpha^*_{21}\frac{\beta_1}{\beta_2} - \alpha^*_{12}\frac{\beta_2}{\beta_1}. \tag{5.3}$$

Therefore, if both agents are benevolent ($\alpha^*_{ij} \in (0,1)$, for $i \neq j$) or if one is benevolent and the other malevolent, then $1 - (\alpha^*_{12})(\alpha^*_{21}) \geq 0$. In this case, $\beta_2 \geq \beta_1$ if and only if $\alpha^*_{21}\beta_1^2 \geq \alpha^*_{12}\beta_2^2$. That is, in order to support equilibrium labor supplies in which each agent contributes its fair share under the MP rule, it must be the case that the more productive agent is also the more altruistic.

One might also enquire of the pattern of mutual concern that prevails in steady state equilibrium. Using the fact that $\beta_2 = 1 - \beta_1$, we plot the values of α^*_{12} and α^*_{21} consistent with (5.3):



Thus, for example, at $\alpha_{12}^* = 0.2$ the steady state values of α_{21}^* as a function of β_1 are depicted in the following figure:



Note, in particular, that for $\beta_1 \in (0.575, 1)$, the steady state value of $\alpha_{21}^* < 0$. Indeed, this pattern is general. For all $\alpha_{12}^* \in [0, 1]$, agent 2 is malevolent for sufficiently large β_1. Furthermore, the range of such β is decreasing in α_{12}^*; that is, when $\alpha_{12}^* = 0$, agent 2 is malevolent for all $\beta_1 \in (0.5, 1)$, and for $\alpha_{12}^* = 1$, the range of such β is $(\frac{\sqrt{2}}{2}, 1)$.

In this case, where agents are rewarded according to their productivities, it is not surprising that 2 is malevolent toward 1 when $\frac{\beta_1}{\beta_2}$ is sufficiently large. However, it is interesting that for given α_{12}^* this tendency is nonmonotonic. That is, while 2 is malevolent for sufficiently large β_1, eventually the extent of malevolence diminishes. In the limit, as $\beta_1 \to 1$ ($\beta_2 \to 0$), 2's malevolence vanishes.

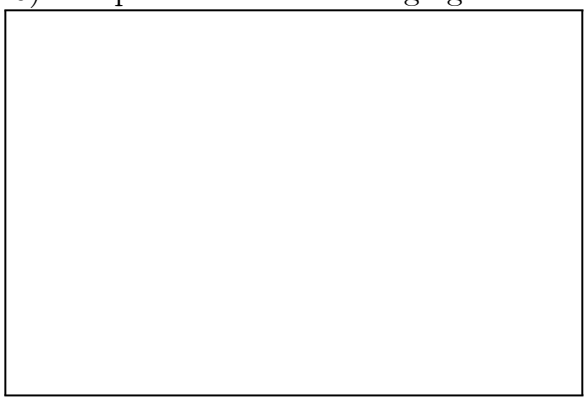
Alternatively, under the egalitarian rule, (5.2) reduces to:

$$\beta_1(1+\alpha_{12}^*) = \beta_2(1+\alpha_{21}^*). \tag{5.4}$$

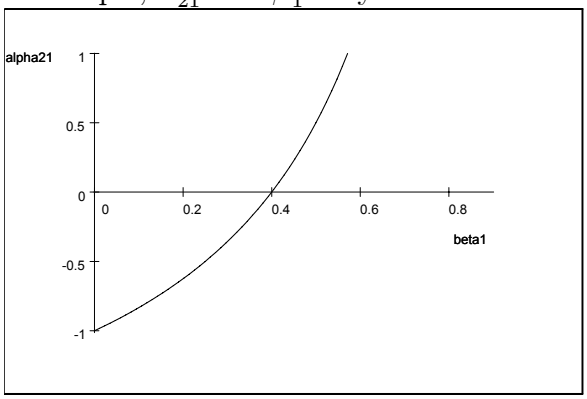
Again substituting $\beta_2 = 1-\beta_1$, we can identify the steady state patterns of concern as follows:

$$\alpha_{21}^* = \frac{\beta_1(2+\alpha_{12}^*)-1}{1-\beta_1}. \tag{5.5}$$

The graph of (5.5) is depicted in the following figure.



At $\alpha_{12}^* = 0.5$, for example, α_{21}^* and β_1 vary as follows:



Here, too, benevolence and malevolence might coexist in a steady state. In this case, however, it is the more productive agent who is malevolent toward the less productive, in accordance with our intuition.

It is interesting to note that under both the MP and the egalitarian allocation rules, the expectation of fair play can generate animosity when such expectations are not fulfilled.

6. Conclusion

In this paper we have presented a model in which agents have extended preferences defined over the welfare of others and such sentiments are determined endogenously on the basis of how agents behave relative to a standard of appropriate behavior. The model allows scope for consideration of the fact that different agents might have different means or abilities and thus may be held to different standards. Also, in general, the model allows for consideration of the role of the individual in determining his or her circumstances.

Unlike similar models involving reciprocity, here agents make no attempt to infer the motives or sentiments of other agents, but rather simply observe their behavior. As such, the model easily allows for third party effects on sentiments; that is, sentiments might change even without direct contact with, or consequence from, another agent.

To demonstrate the potential of such a model in explaining the emergence of different patterns of sentiments, we developed an example involving team production, and we considered three possible standards of behavior. In each case, we identified and/or characterized stationary outcomes, often with surprising results. For example, our results suggest a variety of ways in which benevolent and malevolent agents might stably coexist. Moreover, the pattern of sentiment might differ depending upon the particular standard employed. Also, under the mean standard, the relative industry and pattern of concern in a clustered steady state is ambigous: high productivity agents might work more and be held in greater esteem than low productivity agents or the pattern might be reversed.

We were also able to demonstrate several anomolies. For example, under the transposition norm, it is possible that the agents' mutual degree of concern might increase and yet they would contribute less to the common good. Or, if agents have the same productivities but have different tastes, then even if they receive the same output shares, the unique (interior) stationary outcome might entail mutual malevolence. Also, under the fair standard, it is possible that the degree of malevolence might decline as the disparity between agents productivities increases.

The model thus appears to provide an explanation for a wide variety of behaviors and sentiments among the same set of individuals and to yield a rich and diverse set of predictions.

Within the context of our example, various extensions are possible. First, we have considered only the case in which agents use the same standard to evaluate

the behavior others. In subsequent work we intend to study the interactions among agents who employ different standards. Also, an important direction that we have thus far overlooked is that agents might take into consideration the effect of their actions on other agents' sentiments and hence actions, and they might choose to behave in such a way as to influence them.

Beyond the example presented here, the methodology is quite general and might apply to any case in which behavior is endogenous. An important consequence of this framework is that it profoundly changes the relationship between the individuals in society and the institutional environment.[29] Whereas economists typically take the institutional structure as given and analyze agents' behavior therein, to the extent that institutions influence behavior and behavior affects the sentiments of the agents, the institutions themselves might affect the social composition. This allows scope for considering such compositional effects in the design of social policy. For example, it is possible that some policies might lead to more social cohesion and others to fragmentation or disenfranchisement. In Esteban and Kranich [7], we have begun to explore one such investigation in the context of redistributive taxation. But the potential applications are vast.

References

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

[2] Bolton, G. and A. Ockenfels (2000), "ERC: A Theory of Equity, Reciprocity and Competition," *American Economic Review* **90**, 166-193.

[3] Bowles, S. (1998), "Endogenous Preferences: The Cultural Consequences of Markets and Other Economic Institutions," *Journal of Economic Literature* **36**, 75-111.

[4] Bowles, S. and H. Gintis (1998), "The Evolution of Strong Reciprocity," mimeo, Santa Fe Institute.

[5] Charness, G. and E. Haruvy (2002), "Altruism, Equity and Reciprocity in a Gift-Exchange Experiment: An Encompassing Approach," *Games and Economic Behavior* **40**, 203–231.

[29] On the need to consider such effects see Bowles [3].

[6] Dufwenberg M. and G. Kirchsteiger (1998), "A Theory of Sequential Reciprocity," mimeo, CentER, Tilburg University.

[7] Esteban, J. and L. Kranich (2002), "Redistributive Taxation with Endogenous Sentiments," mimeo.

[8] Falk A. and U. Fischbacher (1998), "A Theory of Reciprocity," mimeo, University of Zurich.

[9] Fehr, E. and S. Gächter (2000), "Fairness and Retaliation: The Economics of Reciprocity," *Journal of Economic Perspectives* **14**, 159-181.

[10] Fehr, E. and K. M. Schmidt (1999), "A Theory of Fairness, Competition and Cooperation," *Quarterly Journal of Economics* **114**, 817-868.

[11] Geanakoplos, J., Pearce, D. and E. Stacchetti (1989), "Psychological Games and Sequential Rationality," *Games and Economic Behavior* **1**, 60-79.

[12] Gilboa, I. and D. Schmeidler (1988), "Information Dependent Games: Can Common Sense be Common Knowledge?" *Economics Letters* **27**, 215-221.

[13] Levine, D. (1998), "Modeling Altruism and Spitefulness in Experiments," *Review of Economic Dynamics* **1**, 593-622.

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

[15] Sethi, R. and E. Somanathan (2001), "Preference Evolution and Reciprocity," *Journal of Economic Theory* **97**, 273-297.

[16] Stark, O. and I. Falk (1998), "Transfers, Empathy Formation, and Reverse Transfers," *AER Papers and Proceedings* **88**, 271-276.