国際大学
INTERNATIONAL UNIVERSITY OF JAPAN

Economics & Management Series *EMS-2022-05*

Belief in Egalitarianism and Meritocracy

Hideaki Goto
International University of Japan

August 2022

IUJ Research Institute
International University of Japan

Belief in Egalitarianism and Meritocracy

Hideaki Goto*

Abstract

Why do people often distribute joint surplus in an egalitarian way even when the payoffs for more productive people are lower than those distributed in a meritocratic way? In particular, does a stationary state exist in which more productive people believe in egalitarianism even when distaste for meritocracy decreases as meritocratic payoffs increase? We extend the Bisin–Verdier model of cultural transmission to address these questions and demonstrate that such a stationary state exists, but is stable only under certain conditions. Therefore, the fractions of people believing in egalitarianism and meritocracy may continue to fluctuate.

JEL Classification: D30, Z13, D63, D91

Keywords: Belief in distributive principles; Egalitarianism; Meritocracy; Cultural transmission.

*International University of Japan, 777 Kokusai-cho, Minami Uonuma-shi, Niigata 949-7277, Japan. Email: h-goto@iuj.ac.jp.

1 Introduction

Distribution rules of jointly produced surplus have been extensively studied (Moulin, 1987, 1988, 2003; Roemer and Silvestre, 1993; Roemer, 1996), among which egalitarian and meritocratic distributions are most frequently considered. However, how egalitarianism can be a distributive norm even for highly productive people, whose egalitarian payoffs are lower than their meritocratic payoffs, remains unclear.

On the one hand, the usual answer to the above question is that egalitarian division is common because it is considered fair by many people. On the other hand, Young (1993) formally illustrates that under certain conditions, 50-50 is the unique stochastically stable division of the evolutionary process in which a (discrete) Nash demand game is played each period by players who learn adaptively with limited memory (see also Young, 1998). The first answer ignores the dilemma between fairness and individual utility: Is material payoff irrelevant to more productive people concerned with fairness? The second approach is credited for providing a sound explanation of why exactly 50-50 division can be a distributive norm. However, it assumes that people's beliefs about fairness do not affect the formation of distributive norms.

By extending the seminal model developed by Bisin and Verdier (2001), in this paper, we incorporate both of the above aspects and analyze the case in which people not only are influenced by methods of distribution proposed by others but also influence the prevalence of distributive principles by their beliefs that surplus *should* be distributed in either an egalitarian or a meritocratic way. In particular, to incorporate the dilemma between fairness and material payoff, we assume that relative payoffs from egalitarian and meritocratic divisions also affect the prevalence of those distributive principles.

More precisely, we assume that individuals, with either low or high productivity, attempt to maintain the current belief in either egalitarianism or meritocracy. If they fail to do so, then they acquire the belief of a randomly drawn individual from the population with the same productivity. Further, extending the Bisin–Verdier model, we assume that the individual's incentive to maintain the current belief, or distaste for a different belief, is weakened as the relative payoff from holding the other belief increases. We demonstrate the existence of a stationary state where individuals with each productivity hold different beliefs. However, such a state is locally stable only under certain conditions. The states where all identically productive individuals belong to the same coalition are shown to be unstable. Therefore, if those conditions are not satisfied, the fraction of individuals with each productivity who believe in each belief may continue to fluctuate forever.

In their paper, Bisin and Verdier (2001) admit that "many aspects related to cultural transmission have been left out of the analysis" and suggest extensions that enable us to consider

"situations in which agents interact in socio-economic environments" or "traits which affect, in a relevant manner, the economic environment the agents face." Our analysis can be regarded as an extension that embeds their model into a particular socio-economic context. Further, by considering the conditions for the cultural emergence of distributive principles in a society, our study provides a positive analysis of the current intellectual debate on the benefits and costs of a meritocratic society (Sandel, 2020; American Journal of Law and Equality, 2021).

The rest of the paper is organized as follows. Section 2 presents the Bisin–Verdier model embedded in the context of our analysis. Section 3 extends the Bisin–Verdier model to incorporate the effect of relative payoffs on the individual's incentive to maintain the current belief. Section 4 concludes the paper and discusses possible further extensions of our extended model.

2 Basic Model

Consider a society that has two observable types of individuals: those with low and high productivity. Let λ_L and λ_H denote the productivity of the former and the latter, respectively, where $0 < \lambda_L < \lambda_H$. Each productivity has a continuum of individuals. N_L and N_H, which are held constant over time, are the measures of low-productivity and high-productivity individuals, respectively.[1]

2.1 Belief in distributive principles and associated payoffs

In addition to productivity, individuals also differ in their beliefs about how surplus should be distributed. We consider two major principles to distribute the total surplus generated by joint production: egalitarianism and meritocracy. In the former, the surplus is equally distributed to all the members of a coalition, whereas in the latter, it is distributed in proportion to individual productivity. Individuals believe in either egalitarianism or meritocracy, join a coalition of individuals who believe in the same principle, and engage in joint production. Thus, at most, two coalitions can exist. A coalition consisting of individuals believing in egalitarianism (meritocracy) is called an *egalitarian (meritocratic) coalition*. Similarly, the payoff to the members in an egalitarian (meritocratic) coalition is called an *egalitarian (meritocratic) payoff*.

The total surplus of each coalition is generated according to a linear production function $f(\Lambda) = \Lambda$, where $\Lambda \equiv n_L\lambda_L + n_H\lambda_H$ is the sum of individual productivities. n_L and n_H are the numbers of low-productivity and high-productivity individuals, respectively.

[1] In the following, the *measure* of individuals will be referred to as the "number" of them for the sake of intuition. Subscripts indicate individual productivities (L for low and H for high) and superscripts represent distributive principles (e for egalitarianism and m for meritocracy).

Let n_L^e and n_L^m (n_H^e and n_H^m) denote the numbers of low-productivity (high-productivity) individuals in the egalitarian and meritocratic coalition, respectively, where $N_L = n_L^e + n_L^m$ and $N_H = n_H^e + n_H^m$. Let θ_k^i be the fraction of individuals with productivity k ($\in \{L, H\}$) in coalition i ($\in \{e, m\}$), that is, $\theta_k^i \equiv n_k^i / N_k$. Then, given λ_k and N_k, the total population and the aggregate productivity in coalition i are respectively

$$N^i(\theta_L^i, \theta_H^i) = n_L^i + n_H^i = \theta_L^i N_L + \theta_H^i N_H$$

and

$$\Lambda^i(\theta_L^i, \theta_H^i) = n_L^i \lambda_L + n_H^i \lambda_H = \theta_L^i N_L \lambda_L + \theta_H^i N_H \lambda_H.$$

As $\theta_k^j = 1 - \theta_k^i$ for $i \neq j$ ($\in \{e, m\}$), N^i and Λ^i are the functions of only θ_L^e and θ_H^e (or θ_L^m and θ_H^m). Let (θ_L^e, θ_H^e) be the state of our system. A state at which each coalition has members with both productivities, that is, (θ_L^e, θ_H^e) with $0 < \theta_L^e < 1$ and $0 < \theta_H^e < 1$, is called *heterogeneous*.

The egalitarian and meritocratic payoffs are given by

$$\text{Egalitarian payoff:} \quad u_k^e(\theta_L^e, \theta_H^e) = \frac{f\left(\Lambda^e(\theta_L^e, \theta_H^e)\right)}{N^e(\theta_L^e, \theta_H^e)} = \frac{\theta_L^e N_L \lambda_L + \theta_H^e N_H \lambda_H}{\theta_L^e N_L + \theta_H^e N_H} \tag{1}$$

$$\text{Meritocratic payoff:} \quad u_k^m(\theta_L^m, \theta_H^m) = \frac{\lambda_k}{\Lambda^m(\theta_L^m, \theta_H^m)} f\left(\Lambda^m(\theta_L^m, \theta_H^m)\right) = \lambda_k \tag{2}$$

for $k \in \{L, H\}$. Obviously, the egalitarian payoff is the same for low- and high-productivity individuals, $u_L^e = u_H^e = u^e$. In addition, it is always (weakly) higher than λ_L and (weakly) lower than λ_H:

$$\lambda_L \leq u^e(\theta_L^e, \theta_H^e) \leq \lambda_H.$$

Moreover, the egalitarian payoff weakly decreases (increases) as low-productivity (high-productivity) individuals join the egalitarian coalition, with equality when the coalition has no high-productivity (low-productivity) individual:

$$\frac{\partial u^e}{\partial \theta_L^e} = \frac{\theta_H^e N_L N_H (\lambda_L - \lambda_H)}{(\theta_L^e N_L + \theta_H^e N_H)^2} \leq 0; \quad \frac{\partial u^e}{\partial \theta_H^e} = \frac{\theta_L^e N_L N_H (\lambda_H - \lambda_L)}{(\theta_L^e N_L + \theta_H^e N_H)^2} \geq 0. \tag{3}$$

2.2 Changes in the belief in distributive principles

We apply the model of Bisin and Verdier (2001) to consider how individuals' beliefs in distributive principles change. Given their own productivity and the share of individuals with each productivity believing in each distributive principle, each individual seeks to maintain their current belief ("direct socialization" in the cultural transmission literature), which succeeds with probability d_k^i ($k \in \{L, H\}$ and $i \in \{e, m\}$). We call this probability *inertia probability*,

reflecting our context. If "direct socialization" fails, which occurs with probability $1 - d_k^i$, the individual's belief is randomly chosen from the population of individuals with the same productivity as their own ("oblique socialization").[2] Let P_k^{ij} denote the probability that the individual with belief i and productivity k comes to believe in principle j. Then for $i \neq j$ we have

$$P_k^{ii} = d_k^i + (1 - d_k^i)\theta_k^i \tag{4}$$

and

$$P_k^{ij} = (1 - d_k^i)(1 - \theta_k^i). \tag{5}$$

Thus, the dynamics of the fraction of individuals with productivity k and belief i in continuous time is characterized by the following equations for $k \in \{L, H\}$ and $i \in \{e, m\}$:

$$\dot{\theta}_k^i = \theta_k^i(1 - \theta_k^i)(d_k^i - d_k^j). \tag{6}$$

The following result immediately follows from equation (6):

Proposition 1. *Suppose that inertia probabilities, d_k^i ($k \in \{L, H\}$ and $i \in \{e, m\}$), are exogenously given. Then $(\theta_L^e, \theta_H^e) \in \{(0,0), (0,1), (1,0), (1,1)\}$ are the stationary states of (6). Further, $\theta_k^e \to 1$ for any $\theta_k^e \in (0, 1]$ if $d_k^e > d_k^m$, and $\theta_k^e \to 0$ for any $\theta_k^e \in [0, 1)$ if $d_k^e < d_k^m$.*

Therefore, egalitarian and meritocratic coalitions may coexist with exogenous inertia probabilities. For example, if $d_L^e > d_L^m$ and $d_H^e < d_H^m$, then the only asymptotically stable state is that in which all low-productivity (high-productivity) individuals belong to an egalitarian (meritocratic) coalition and receive $u^e(1, 0) = \lambda_L$ ($u_H^m = \lambda_H$). At the states $(0, 0), (0, 1)$, and $(1, 0)$, individuals' payoffs equal their productivities, while at the state $(1, 1)$, low-productivity (high-productivity) individuals receive a higher (lower) payoff than their productivity.

2.3 Endogenous inertia probability with fixed distastes against different principles

We next consider the case where inertia probabilities are endogenously determined. Let V^{ij} be the utility, evaluated at the time when the individual believes in principle i, when the belief

[2] Alternatively, we could consider the more usual scenario where parents attempt to instill in their children their own belief. We have the same results as those obtained following the scenario presented in the main text if we assume that the productivity of a child is always the same as the parent's, which is not realistic. Bénabou and Tirole (2006) consider two similar scenarios in their analysis of belief in a just world, one in which parents attempt to direct their children to a positive belief and the other in which individuals manipulate their own belief.

changes from i to $j(\neq i)$. The individual's utility is V^{ii} if the belief does not change.[3] Then, $V^{ii} - V^{ij}$ can be regarded as *distaste* for a different principle. In this subsection, we follow Bisin and Verdier (2001) and assume that V^{ii} and V^{ij} are fixed with $V^{ii} > V^{ij}$.

Given V^{ii}, V^{ij}, (4), and (5), each individual chooses the inertia probability d_k^i that maximizes their expected utility:

$$\max_{d_k^i} \sum_j P_k^{ij} V^{ij} - C(d_k^i), \tag{7}$$

where C is a cost function representing costs to maintain the same belief. For analytical tractability, we follow Bisin et al. (2009) and Montgomery (2010) and assume a quadratic cost function, $C(d) = d^2/2$.[4]

The first-order condition is

$$d_k^i(\theta_k^i, \Delta^{ij}) = (1 - \theta_k^i)\Delta^{ij},$$

where Δ^{ij} is distaste for the other principle: $\Delta^{ij} \equiv V^{ii} - V^{ij}$ $(i \neq j)$. The dynamics of the fraction of individuals with productivity k and belief i is determined by

$$\dot{\theta}_k^i = \theta_k^i(1 - \theta_k^i)\left(d_k^i(\theta_k^i, \Delta^{ij}) - d_k^j(\theta_k^j, \Delta^{ji})\right). \tag{8}$$

From equation (8), $\dot{\theta}_k^i = 0$ when $\theta_k^i = 0, 1$ and $d_k^i = d_k^j$. As $\theta_k^i = 1 - \theta_k^j$, the last equality holds when the following equation is satisfied:

$$(1 - \theta_k^i)\Delta^{ij} = \theta_k^i \Delta^{ji}.$$

Therefore, the fraction of individuals with productivity k who believe in principle i is given by

$$\hat{\theta}_k^i = \frac{\Delta^{ij}}{\Delta^{ij} + \Delta^{ji}}. \tag{9}$$

Proposition 2. *When d_k^i is endogenously determined, $\{(0,0),(0,1),(1,0),(1,1),(\hat{\theta}_L^e, \hat{\theta}_H^e)\}$ are the stationary states of (8). Moreover, $(\theta_L^e, \theta_H^e) \to (\hat{\theta}_L^e, \hat{\theta}_H^e)$ for any $\{(\theta_L^e, \theta_H^e) \mid 0 < \theta_L^e < 1, 0 < \theta_H^e < 1\}$.*

Proof. The above results are obtained by considering $k = L$ and $k = H$ separately and ap-

[3] In Bisin and Verdier (2001), V^{ij} are based on the parent's evaluation of the child's optimal choice from the parent's rather than the child's perspective. As the parent with the same belief as that of the child evaluates the child's choice more highly, $V^{ii} > V^{ij}$ holds. We assume in this subsection that the optimal d_k^i satisfies $0 \leq d_k^i \leq 1$, given the values of V^{ii} and V^{ij}.

[4] The results of the current subsection hold for a more general cost function C with $C', C'' > 0$ and $C(0) = C'(0) = 0$, in which case, the first-order condition is $C'(d_k^i) = (1 - \theta_k^i)\Delta^{ij}$. Thus, $d_k^i = d_k^j$ if and only if $(1 - \theta_k^i)\Delta^{ij} = (1 - \theta_k^j)\Delta^{ji} = \theta_k^i \Delta^{ji}$, as in the main text.

plying the proof of Proposition 1 in Bisin and Verdier (2001) or that found in Section III.A of Montgomery (2010). □

Therefore, a unique, globally stable heterogeneous state exists in this case. The fraction of individuals believing in a certain principle is determined by the relative strength of the distastes for the other distributive principles: if $V^{ii} - V^{ij} > V^{jj} - V^{ji}$, then more individuals believe in principle i than in principle j in the stable stationary state.

The above analysis can be easily extended to the case where distastes differ depending on productivity. If, for example, distaste for egalitarianism is greater and that for meritocracy is lower for high-productivity than for low-productivity individuals, then Δ^{me} is greater and Δ^{em} is smaller for high-productivity individuals, which results in a state with a higher fraction of high-productivity than low-productivity individuals believing in a meritocracy.

In any case, the fraction of individuals with each productivity believing in each principle in the stable stationary state is completely determined by fixed distastes for different distributive principles.

3 Distastes Depending on Relative Payoffs

3.1 The extended model

We now extend the Bisin–Verdier model to analyze the case in which each individual cares not only about the utility from their own belief (or the change thereof) but also about their material payoffs. More precisely, each individual always values their current belief more highly than the other one, but distaste for the other distributive principle is weakened as the payoff they would receive had they held that belief relatively increases.[5] That is, distaste for a different principle $V^{ii} - V^{ij}$, which is assumed to be always positive, changes depending on the individual's payoffs associated with two distributive principles.

Formally, we let $V^{ii} - V^{ij} = D(\Delta u_k^i)$, where Δu_k^i is the difference between the payoffs for an individual with productivity k when the belief is in principle i and when it is in principle j: $\Delta u_k^i \equiv u_k^i - u_k^j$. D is a *distaste function*, which we assume satisfies the following conditions:

Assumption 1. *D is continuously differentiable and $D, D' > 0$ for any Δu_k^i, $D(\lambda_H - \lambda_L) \leq 1$, and $D \to \infty$ as $\Delta u_k^i \to \infty$.*

As in subsection 2.3, the individual solves the maximization problem (7). In contrast to that case, however, given the fraction of individuals with productivity l ($\neq k$) belonging to the

[5]The *would-be* payoff of an individual with productivity k in the egalitarian coalition when no (other) individual with the same productivity belongs to the coalition, $u_k^e(\theta_L^e, \theta_H^e)$ for $\theta_k^e = 0$, is defined as the limit of u_k^e as $\theta_k^e \to 0$, which is simply the productivity level other than that of the individual, λ_l ($l \neq k$).

same coalition i, θ_l^i, the first-order condition for the optimal level of inertia probability is now given by:[6]

$$d_k^i(\theta_k^i;\theta_l^i) = (1-\theta_k^i)D(\Delta u_k^i). \tag{10}$$

The dynamics of the fractions of low- and high-productivity individuals believing in egalitarianism is determined by

$$\dot{\theta}_L^e = \theta_L^e(1-\theta_L^e)F(\theta_L^e,\theta_H^e) \tag{11}$$

and

$$\dot{\theta}_H^e = \theta_H^e(1-\theta_H^e)G(\theta_L^e,\theta_H^e), \tag{12}$$

where

$$F(\theta_L^e,\theta_H^e) = (1-\theta_L^e)D\left(u^e(\theta_L^e,\theta_H^e)-\lambda_L\right) - \theta_L^e D\left(\lambda_L - u^e(\theta_L^e,\theta_H^e)\right)$$

and

$$G(\theta_L^e,\theta_H^e) = (1-\theta_H^e)D\left(u^e(\theta_L^e,\theta_H^e)-\lambda_H\right) - \theta_H^e D\left(\lambda_H - u^e(\theta_L^e,\theta_H^e)\right).$$

Note that

$$\dot{\theta}_L^m = -\dot{\theta}_L^e \quad \text{and} \quad \dot{\theta}_H^m = -\dot{\theta}_H^e. \tag{13}$$

F and G are continuously differentiable (except at (θ_L^e,θ_H^e) such that $\theta_L^e N_L + \theta_H^e N_H = 0$). We have $F(0,\theta_H^e) > 0$ and $F(1,\theta_H^e) < 0$ for any θ_H, and $G(\theta_L^e,0) > 0$ and $G(\theta_L^e,1) < 0$ for any θ_L^e. By differentiating F and G, we also have $\partial F/\partial\theta_L^e < 0$, $\partial F/\partial\theta_H^e \geq 0$ (with equality when $\theta_L^e = 0$), and $\partial G/\partial\theta_L^e \leq 0$ (with equality when $\theta_H^e = 0$). Moreover, differentiating $\dot{\theta}_L^e$ and $\dot{\theta}_H^e$ with respect to θ_L^e and θ_H^e gives

$$\frac{\partial\dot{\theta}_L^e}{\partial\theta_L^e} = (1-2\theta_L^e)F(\theta_L^e,\theta_H^e) + \theta_L^e(1-\theta_L^e)\frac{\partial F(\theta_L^e,\theta_H^e)}{\partial\theta_L^e}, \tag{14}$$

$$\frac{\partial\dot{\theta}_L^e}{\partial\theta_H^e} = \theta_L^e(1-\theta_L^e)\frac{\partial F(\theta_L^e,\theta_H^e)}{\partial\theta_H^e} \geq 0, \tag{15}$$

$$\frac{\partial\dot{\theta}_H^e}{\partial\theta_L^e} = \theta_H^e(1-\theta_H^e)\frac{\partial G(\theta_L^e,\theta_H^e)}{\partial\theta_L^e} \leq 0, \tag{16}$$

and

$$\frac{\partial\dot{\theta}_H^e}{\partial\theta_H^e} = (1-2\theta_H^e)G(\theta_L^e,\theta_H^e) + \theta_H^e(1-\theta_H^e)\frac{\partial G(\theta_L^e,\theta_H^e)}{\partial\theta_H^e}, \tag{17}$$

[6] As $d_k^i \in [0,1]$, this condition shows that D must be positive in the current model. Future work could consider the case where D can take negative values.

where

$$\frac{\partial G}{\partial \theta_H^e} = -D(u^e - \lambda_H) - D(\lambda_H - u^e) + \frac{\partial u^e}{\partial \theta_H^e}\left[(1-\theta_H^e)D'(u^e - \lambda_H) + \theta_H^e D'(\lambda_H - u^e)\right]. \quad (18)$$

From (11) and (12), we find that, in addition to $(\theta_L^e, \theta_H^e) = \{(0,0),(0,1),(1,0),(1,1)\}$, a state $({\theta_L^e}^*, {\theta_H^e}^*)$ such that

$$F({\theta_L^e}^*, {\theta_H^e}^*) = G({\theta_L^e}^*, {\theta_H^e}^*) = 0 \quad (19)$$

is also stationary.

These and the stability characteristics of stationary states are summarized in the following proposition:

Proposition 3. *$(\theta_L^e, \theta_H^e) = \{(0,0),(0,1),(1,0),(1,1)\}$ are all stationary but unstable. Moreover, a heterogeneous stationary state $({\theta_L^e}^*, {\theta_H^e}^*)$ exists where less (more) than half of the high-productivity (low-productivity) individuals belong to the egalitarian coalition: $1/2 < {\theta_L^e}^* < 1$ and $0 < {\theta_H^e}^* < 1/2$. The state $({\theta_L^e}^*, {\theta_H^e}^*)$ is locally stable if $\partial G/\partial \theta_H^e < 0$ holds at $({\theta_L^e}^*, {\theta_H^e}^*)$.*

Proof. (a) *$(\theta_L^e, \theta_H^e) = \{(0,0),(0,1),(1,0),(1,1)\}$ are all stationary but unstable*:

From (11), (12), and (13), these states are clearly stationary. From (14) and (17), we have $\partial \dot{\theta}_k^e/\partial \theta_k^e|_{\theta_k^e=0} > 0$ and $\partial \dot{\theta}_k^e/\partial \theta_k^e|_{\theta_k^e=1} > 0$ for $k \in \{L, H\}$. Thus, $\dot{\theta}_k^e > 0$ ($\dot{\theta}_k^e < 0$) in any arbitrarily small neighborhood of $\theta_k^e = 0$ ($\theta_k^e = 1$). From (13) and $\theta_k^m = 1 - \theta_k^e$, $\dot{\theta}_k^m < 0$ ($\dot{\theta}_k^m > 0$) in any arbitrarily small neighborhood of $\theta_k^m = 1$ ($\theta_k^m = 0$). Therefore, any state where all the individuals with the same productivity belong to the same coalition is unstable.

(b) *Existence of $({\theta_L^e}^*, {\theta_H^e}^*)$*:

We first note that $F(1/2, 0) = 0$. From $\partial F/\partial \theta_L^e < 0$ and $\partial F/\partial \theta_H^e > 0$ (for $\theta_L^e > 0$), F takes the value of zero for $\theta_H^e \geq 0$ only for $\theta_L^e \geq 1/2$. From $F(1/2, 1) > 0$, $F(1,1) < 0$, and the continuity of $F(\theta_L^e, 1)$ in θ_L^e, by the intermediate value theorem, $\theta_L^e \in (1/2, 1)$ exists at which $F(\theta_L^e, 1) = 0$. As $\partial F/\partial \theta_L^e < 0$, such a value is unique. Let θ_L^{max} denote that value. As $\partial F/\partial \theta_L^e < 0$ and $\partial F/\partial \theta_H^e > 0$, for any $\theta_L^e \in [1/2, \theta_L^{max}]$, a unique $\theta_H^e \in [0,1]$ exists, which is denoted by $h(\theta_L^e)$, such that $F(\theta_L^e, h(\theta_L^e)) = 0$. By the implicit function theorem, h is continuously differentiable with $h' > 0$ on $[1/2, \theta_L^{max}]$.[7] Obviously, $h(1/2) = 0$ and $h(\theta_L^{max}) = 1$.

For G, let θ_L^e take any values in $(-\theta_H^e N_H/N_L, \infty)$ for the moment, where $u^e \to \infty$ as $\theta_L^e \to -\theta_H^e N_H/N_L$ and $u^e \to \lambda_L$ as $\theta_L^e \to \infty$. Hence, $u^e - \lambda_H \to \infty$ as $\theta_L^e \to -\theta_H^e N_H/N_L$

[7] If we allow θ_H^e to take values less than 0 and greater than 1, then the theorem applies to (around) $\theta_L^e = 1/2$ and $\theta_L^e = \theta_L^{max}$ as well.

and $u^e - \lambda_H \to \lambda_L - \lambda_H (< 0)$ as $\theta_L^e \to \infty$. Let $\tilde{\theta}_H = D(\lambda_L - \lambda_H)/[D(\lambda_L - \lambda_H) + D(\lambda_H - \lambda_L)]$. Note that $\tilde{\theta}_H$ is less than 1/2. Then, $\lim_{\theta_L^e \to \infty} G(\theta_L^e, \tilde{\theta}_H) = (1 - \tilde{\theta}_H) D(\lambda_L - \lambda_H) - \tilde{\theta}_H D(\lambda_H - \lambda_L) = 0$. Given any $\theta_H^e \in (\tilde{\theta}_H, 1 - \varepsilon)$, for an arbitrarily small $\varepsilon > 0$, $G(\theta_L^e, \theta_H^e) > 0$ as $\theta_L^e \to -\theta_H^e N_H / N_L$ and $G(\theta_L^e, \theta_H^e) < 0$ as $\theta_L^e \to \infty$. As $\partial G / \partial \theta_L^e < 0$ for $\theta_H^e > 0$, a unique θ_L^e, denoted by $g(\theta_H^e)$, exists such that $G(g(\theta_H^e), \theta_H^e) = 0$. By the implicit function theorem, g is continuously differentiable on $(\tilde{\theta}_H, 1 - \varepsilon)$, $g \to \infty$ as $\theta_H^e \to \tilde{\theta}_H$ and $g \to -\theta_H^e N_H / N_L$ as $\theta_H^e \to 1$. In particular, from $G(0, 1/2) = 0$, we have $g(1/2) = 0$.

As both h and g are continuous, a heterogeneous state $({\theta_L^e}^*, {\theta_H^e}^*)$ with $1/2 < {\theta_L^e}^* < \theta_L^{max}$ and $\tilde{\theta}_H < {\theta_H^e}^* < 1/2$ exists at which equation (19) holds.

(c) *Stability of* $({\theta_L^e}^*, {\theta_H^e}^*)$:

The Jacobian matrix of the system is given by

$$J = \begin{bmatrix} \frac{\partial \dot{\theta}_L^e}{\partial \theta_L^e} & \frac{\partial \dot{\theta}_L^e}{\partial \theta_H^e} \\ \frac{\partial \dot{\theta}_H^e}{\partial \theta_L^e} & \frac{\partial \dot{\theta}_H^e}{\partial \theta_H^e} \end{bmatrix} \tag{20}$$

evaluated at (θ_L^e, θ_H^e). If its trace is negative and determinant is positive at $({\theta_L^e}^*, {\theta_H^e}^*)$, then the state is locally stable. From (14), (15), (16), and $F({\theta_L^e}^*, {\theta_H^e}^*) = G({\theta_L^e}^*, {\theta_H^e}^*) = 0$, we have $\partial \dot{\theta}_L^e / \partial \theta_L^e < 0$, $\partial \dot{\theta}_L^e / \partial \theta_H^e > 0$, and $\partial \dot{\theta}_H^e / \partial \theta_L^e < 0$ at $({\theta_L^e}^*, {\theta_H^e}^*)$. From (17) and $G({\theta_L^e}^*, {\theta_H^e}^*) = 0$, $\partial \dot{\theta}_H^e / \partial \theta_H^e < 0$ if and only if $\partial G / \partial \theta_H^e < 0$ (both at $({\theta_L^e}^*, {\theta_H^e}^*)$). Therefore, $({\theta_L^e}^*, {\theta_H^e}^*)$ is locally stable if $\partial G / \partial \theta_H^e < 0$ at the state.

□

Note that in the Bisin–Verdier model in subsection 2.3, the inertia probability, d_k^i, satisfies the *cultural substitution* property (Bisin and Verdier, 2001), that is, it is continuous and strictly decreasing in θ_k^i with $d_k^i = 0$ for $\theta_k^i = 1$, whereas in the extended model, the inertia probability for high-productivity individuals, d_H^i, may or may not be decreasing in θ_H^i. On the one hand, high-productivity individuals' effort to maintain their current belief decreases as more high-productivity individuals hold the same belief. On the other hand, their (relative) payoff increases as more of them start belonging to the same coalition, which raises their distaste for the other principle as well as their optimal effort level. These two opposing effects lead to the potential instability of the heterogeneous stationary states of the extended model.

The corollary below provides sufficient conditions for the local stability of $({\theta_L^e}^*, {\theta_H^e}^*)$:

Corollary 1. *If* $D' < D$, $\lambda_H - \lambda_L \leq 1$, *and* $N_L \geq 2N_H$, *then the heterogeneous stationary state* $({\theta_L^e}^*, {\theta_H^e}^*)$ *is locally stable.*

Proof. By differentiating $\partial u^e/\partial \theta_H^e$ (given by (3)) with respect to θ_L^e, we observe that $\partial u^e/\partial \theta_H^e$ is decreasing in θ_L^e if and only if $\theta_H^e N_H < \theta_L^e N_L$, which is satisfied for ${\theta_L^e}^* > 1/2$ and ${\theta_H^e}^* < 1/2$ if $N_L \geq N_H$. Equation (3) clarifies that $\partial u^e/\partial \theta_H^e$ is decreasing in θ_H^e. Thus, $\partial u^e/\partial \theta_H^e < 1$ at $({\theta_L^e}^*, {\theta_H^e}^*)$ if $N_L \geq N_H$ and $\partial u^e/\partial \theta_H^e \leq 1$ holds at $(\theta_L^e, \theta_H^e) = (1/2, 0)$, which is the case if $N_L \geq 2N_H$ and $\lambda_H - \lambda_L \leq 1$. Therefore, from (18), $\partial G/\partial \theta_H^e < 0$ at $({\theta_L^e}^*, {\theta_H^e}^*)$ if $D' < D$, $\lambda_H - \lambda_L \leq 1$, and $N_L \geq 2N_H$. □

Under the above conditions, the effect of "oblique socialization" on the inertia probability outweighs the effect of relative payoffs on that probability. Consequently, the "cultural substitution" property is regained, at least locally, and thus, the local stability of the heterogeneous stationary state is obtained.

4 Conclusion

By extending the seminal model of Bisin and Verdier (2001), we have demonstrated that a heterogeneous stationary state exists even when individuals consider payoffs associated with each belief when deciding how much they need to strive to retain their current beliefs. At such a state, less (more) than half of the high-productivity (low-productivity) individuals belong to the egalitarian coalition. However, the stationary state is not always stable. The potential instability of the heterogeneous stationary state arises from the fact that the inertia probability of high-productivity individuals with a certain belief may or may not be decreasing in the fraction of individuals with the same belief and productivity. As the fraction of high-productivity individuals believing in the same belief increases, the individual with the same productivity needs to expend less effort to maintain the same belief. However, as more high-productivity individuals start belonging to the same coalition, the (relative) payoff from holding the belief increases, which leads the individual to try harder to maintain the current belief. We have provided sufficient conditions under which the heterogeneous stationary state is locally stable. As any state at which all the individuals with the same productivity belong to the same coalition is unstable, the state may continue to fluctuate if those conditions are not met.

This study is an attempt to embed the Bisin–Verdier model into a specific socio-economic context. However, more work needs to be conducted for improvement. One could examine cases with more distributive principles than egalitarianism and meritocracy and/or with more than two levels of individual productivity. Further, if we allow individuals with the same belief and different productivities to form different coalitions and the production function to be increasing returns, then finding the resulting coalitional structures might be a worthwhile pursuit (e.g., Farrell and Scotchmer, 1988; Herings et al., 2021). Our analysis indicates that the

expected results are likely to be in a sharp contrast to the cases in which people have no belief in distribution principles and move from one coalition to another based only on their material payoffs. Such further extensions may help explain complex coalitional structures in a variety of fields including politics, economics, and everyday life.

References

American Journal of Law and Equality, 2021. Symposium on Michael Sandel's The Tyranny of Merit Volume 1.

Bénabou, R., Tirole, J., 2006. Belief in a just world and redistributive politics. Quarterly Journal of Economics 121, 699–746.

Bisin, A., Topa, G., Verdier, T., 2009. Cultural transmission, socialization and the population dynamics of multiple-trait distributions. International Journal of Economic Theory 5, 139–154.

Bisin, A., Verdier, T., 2001. The economics of cultural transmission and the dynamics of preferences. Journal of Economic Theory 97, 298–319.

Farrell, J., Scotchmer, S., 1988. Partnerships. Quarterly Journal of Economics 103, 279–297.

Herings, P.J.J., Saulle, R.D., Seel, C., 2021. The last will be first, and the first last: Segregation in societies with relative pay-off concerns. The Economic Journal 131, 2119–2143.

Montgomery, J.D., 2010. Intergenerational cultural transmission as an evolutionary game. American Economic Journal: Microeconomics 2, 115–136.

Moulin, H., 1987. Equal or proportional division of a surplus, and other methods. International Journal of Game Theory 16, 161–186.

Moulin, H., 1988. Axioms of Cooperative Decision Making. Cambridge University Press, Cambridge.

Moulin, H., 2003. Fair Division and Collective Welfare. MIT Press, Cambridge.

Roemer, J.E., 1996. Theories of Distributive Justice. Harvard University Press, Cambridge.

Roemer, J.E., Silvestre, J., 1993. The proportional solution for economies with both private and public ownership. Journal of Economic Theory 59, 426–444.

Sandel, M.J., 2020. The Tyranny of Merit: What ' s Become of the Common Good? Farrar, Straus and Giroux, New York.

Young, H.P., 1993. An evolutionary model of bargaining. Journal of Economic Theory 59, 145–168.

Young, H.P., 1998. Individual Strategy and Social Structure: An Evolutionary Theory of Institutions. Princeton University Press, New Jersey.