ACCOUNTING FOR THE LONG-TERM STABILITY OF THE WELFARE-STATE REGIMES IN A MODEL WITH DISTRIBUTIVE PREFERENCES AND SOCIAL NORMS

Gilles Le Garrec

SCIENCES PO OFCE WORKING PAPER n° 01/2023

EDITORIAL BOARD

CONTACT US

OFCE
10 place de Catalogne | 75014 Paris | France
Tél. +33 1 44 18 54 24
www.ofce.fr

WORKING PAPER CITATION

This Working Paper:
Gilles Le Garrec
Accounting for the long-term stability of the welfare-state regimes in a model with distributive preferences and social norms
Sciences Po OFCE Working Paper, n°01/2023.
Downloaded from URL: www.ofce.sciences-po.fr/pdf/dtravail/OFCEWP2023-01.pdf
DOI - ISSN

ABOUT THE AUTHORS

Gilles Le Garrec, Sciences Po-OFCE.
Email Address: gilles.legarrec@sciencespo.fr

ABSTRACT

After the Esping-Andersen' (1990) seminal study, welfare states are standardingly clustered in three identifiable regimes, liberal for Anglo-Saxon countries, corporatist for Continental Europe and social-democratic for Nordic countries, into which the levels of income redistribution can be ranked, from the lowest for the first to the highest for the last. By finding that most European continental countries are now clustered in the high-taxation group along with Nordic countries, a recent study by Péligry and Ragot (2022) has suggested that the welfare states can evolve and change over time, casting doubt on the long-term stability of the canonical clustering. To study this issue, focusing on the redistributive features of welfare states, we develop an overlapping generations model with distributive preferences and social norms, in which we assimilate a welfare-state regime to a stable stationary state with perfect stable expectations. Three configurations with three regimes may arise depending on how unfair the income distribution is perceived (determined by luck rather than effort). In the one associated with the least unfairness, only the low-redistribution regime is truly stable. The two others, while responding to the definition of a stable stationary state with perfect expectations, can be destabilized by a self-fulfilling belief. They are only locally determinate. However, the clustering in three regimes is supposed to persist over time, even if the sets of countries composing the intermediate and the high-redistribution regimes can change. In the configuration associated with intermediate unfairness, only the intermediate-redistribution regime remains only locally determinate. This configuration predicts that countries with an intermediate welfare regime will end (in an indeterminate time) with a low-redistribution one, and the final clustering will encounter only two regimes, the low- and the high-redistribution. Finally, it is only in the configuration associated with the greatest unfairness that the clustering of the welfare states in three regimes is truly stable, with the sets of countries composing the different models unchanged over time.

KEYWORDS: Redistribution, voting behavior, fairness, endogenous preferences.

JEL: H53, D72, D64.

Accounting for the long-term stability of the welfare-state regimes in a model with distributive preferences and social norms

Gilles Le Garrec*

Sciences Po - OFCE

Abstract

After the Esping-Andersen' (1990) seminal study, welfare states are standardingly clustered in three identifiable regimes, liberal for Anglo-Saxon countries, corporatist for Continental Europe and social-democratic for Nordic countries, into which the levels of income redistribution can be ranked, from the lowest for the first to the highest for the last. By finding that most European continental countries are now clustered in the high-taxation group along with Nordic countries, a recent study by Péligry and Ragot (2022) has suggested that the welfare states can evolve and change over time, casting doubt on the long-term stability of the canonical clustering. To study this issue, focusing on the redistributive features of welfare states, we develop an overlapping generations model with distributive preferences and social norms, in which we assimilate a welfare-state regime to a stable stationary state with perfect stable expectations. Three configurations with three regimes may arise depending on how unfair the income distribution is perceived (determined by luck rather than effort). In the one associated with the least unfairness, only the low-redistribution regime is truly stable. The two others, while responding to the definition of a stable stationary state with perfect expectations, can be destabilized by a self-fulfilling belief. They are only locally determinate. However, the

*gilles.legarrec@sciencespo.fr

Without implication for any responsability, I would like to thank Jérôme Creel, Frédéric Gannon, Xavier Ragot and Vincent Touzé for their comments and suggestions, as well as the participants of the EWPM Workshop in Paris 2022.

clustering in three regimes is supposed to persist over time, even if the sets of countries composing the intermediate and the high-redistribution regimes can change. In the configuration associated with intermediate unfairness, only the intermediate-redistribution regime remains only locally determinate. This configuration predicts that countries with an intermediate welfare regime will end (in an indeterminate time) with a low-redistribution one, and the final clustering will encounter only two regimes, the low- and the high-redistribution. Finally, it is only in the configuration associated with the greatest unfairness that the clustering of the welfare states in three regimes is truly stable, with the sets of countries composing the different models unchanged over time.

Keywords: redistribution, voting behavior, fairness, endogenous preferences
JEL: H53, D72, D64

1 Introduction

As a set of institutions and redistributive policies aiming at protecting the citizens from undesirable market outcomes and promoting fairness in the distribution of wealth, welfare states are standardingly clustered in three identifiable regimes according to the Esping-Andersen' (1990) seminal study. In the liberal welfare regime, archetype of Anglo-Saxon countries, support is targeted at the poor, those unable to generate adequate income in the labor market. Flat-rate benefits are low and available to all who meet the eligibility criteria. In contrast, in the social-democratic welfare regime, archetype of Nordic countries, benefits are universal (available to all citizens) and set at a fairly high level. Accordingly, social-democratic welfare states engage in much more income redistribution than do liberal welfare states. Finally, in the corporatist welfare regime, archetype of Continental Europe, benefits are also high but linked to the contributions jointly paid by the workers and their employers. Therefore, this regime can be classified as intermediate in terms of income redistribution. As the canonical clustering of welfare-state regimes in social sciences, it has been extensively debated, criticized, extended. Criticized, for example, for its willingness to put too many institutional and cultural specificities in too few categories[1], the clustering of the welfare states in three regimes has received some support, for the year 2018, in a recent statistical study by Péligry and Ragot (2022). In line with Esping-Andersen (1990), they find that the low-taxation group is composed mostly of Anglo-Saxon countries as the United-States, the UK, Ireland, New-Zealand and Canada. However, they also show that most European continental countries (France, Italy, Germany, Belgium, Austria) that are associated with the corporatist regime in Esping-Andersen (1990) are clustered in the high-taxation group, along with Sweden, Denmark and Finland (interestingly, Norway belongs to the intermediate regime, not to the social-democratic). This finding may suggest that the welfare states of these countries have evolved over time to become currently closer to the high-redistribution welfare-state regime. It thus raises the long-term stability of the welfare-state regimes, and especially of the intermediate one. This article, focusing on the ability of welfare states to redistribute income, precisely tackles this issue.

Following robust empirical evidence that fairness and culture are two important components of redistributive attitudes, in this article we develop an overlapping generations model

[1] Additional welfare-state regimes have been proposed to represent Southern European countries (Ferrera, 1996) and East Asian countries (Goodman and Peng, 1996), which would lead to five institutional regimes consistent with the five types of capitalism described by Amable (2005). It should be noted that the level of detail provided in this literature to the description of the institutions goes well beyond the scope of this article.

that merges the fairness approach of Angeletos and Alesina (2005) and the mechanism for the cultural transmission of the strength of the moral norm proposed by Le Garrec (2018). It results in a dynamics of redistribution that is both backward- and forward-looking, i.e. both history- and belief-dependent, in which we assimilate a welfare-state regime to a stable stationary state with perfect stable expectations. We then show that, under some conditions, three configurations with three welfare-state regimes arise according to the degree of unfairness in the income distribution measured by the ratio of the variance of an idiosyncratic shock to the variance of individual talents (this reflects how important luck is in determining income, where luck is an unfair component of income).

In the first configuration, associated with the least unfair income distribution, only the low-redistribution regime is truly stable. The two others, while responding to the definition of a stationary state with perfect expectations, can be destabilized by a self-fulfilling belief. They are only stable in local-belief, or, in other words, locally but not globally determinate. To summarize, we will say they are weakly stable. Starting from the intermediate regime, the tax rate can jump up then converge to the one of the high-redistribution regime. That may help explain why countries formerly associated with the intermediate regime have become now closer to the high-redistribution regime. However, it does not mean that the clustering of institutions would end with only two regimes, the low- and the high-redistribution ones. Indeed, if the intermediate regime can converge towards the high-redistribution one, the reverse is also possible. The clustering in three regimes appears stable in the long term in this configuration, even if the sets of countries composing the intermediate and the high-redistribution regimes can change. This characteristic underlines the cultural similarity (shared values) of the two latter regimes, both of them implementing an income redistribution being closer to fair level than in the low-redistribution regime because citizen within exhibit a stronger (endogenous) moral universalism characterized by aversion for unfairness.

In the second configuration, associated with intermediate unfairness in the income distribution, only the intermediate-redistribution regime remains only locally determinate. It has important consequences in terms of long-term stability of the three-regime clustering and the existence of the intermediate welfare-state regime. Indeed, in a finite time, we can figure out that the intermediate regime could be destabilized with self-fulfilling beliefs sustaining greater redistribution. If it happens, redistribution becomes effectively greater. However, contrary to the previous configuration, the jump up is too small to initiate a process of convergence towards the high-redistribution regime. The leading phenomenum is that the implemented redistributive institution is too small compared to the one that is perceived as fair. As a consequence,

consistently with the socialization process proposed by Le Garrec (2018), the generation that is young in this period is socialized in an environement whose practices and institutions are far from reflecting fairness, and internalization of the observed norm *you should behave according to your own interest* reduces the concern for fairness. A process of convergence towards the low-redistribution regime is initiated. This configuration predicts then that, in a finite time, countries with an intermediate welfare-state regime will end with a low-redistribution one, and the final clustering will encounter only two regimes, the low- and the high-redistribution, which cannot be destibilized by other beliefs, i.e. which are globally determined.

Finally, in the third configuration, associated with the greatest unfairness in the income distribution, the three regimes of welfare state are truly stable. More specifically, the sets of countries composing the different regimes are unchanged over time. In the intermediate regime, compared to the previous configuration, individuals are no more sufficiently close to a fair institution for believing that higher redistribution is an option. The strength of the concern for fairness is too low. This creates determinacy, as in the low-redistribution regime. It is the reverse of the high-redistribution regime in which it is the proximity with the fair institution that prevents lower redistribution to be a credible alternative. The strength of the concern for fairness is too high for that. Given the rise in inequalities observed in recent years, it is possible that, after a period of reallocation of the European countries with high taxation between the intermediate and social-democratic regimes, a truly stable clustering of the welfare states in three categories will appear.

This article is related to two strands of literature. First, it builds on the literature linking redistribution to other-regarding motives. In a departure from traditional economics, a large body of experimental evidence (see Fehr and Schmidt, 2006, for an overview) shows that individuals do not behave always selfishly as assumed, and that other-regarding motives matter in particular in explaining redistributive attitudes (Tyran and Sausgruber, 2006; Ackert et al., 2007; Schildberg-Hörisch, 2010; Durante et al., 2014; Rustichini and Vostroknutov, 2014; Lefgren et al., 2016; Kerschbamer and Müller, 2020). For these unselfish motives to translate into a desire for greater national redistribution, Enke et al. (2022) nevertheless specify that they must reflect universal moral, i.e. impersonal principles that are not tied to particular groups, family, friends or any socially-related individuals. It is not that universalist individuals are necessarely more generous, but, as underlined by Enke et al. (2022), the willingness to help his neighbor does not predict support for redistribution, or even the reverse. Among the universal moral values, fairness appears to play a key role in explaining redistributive policies. First,

when studying redistributive attitudes, surveys clearly show that individuals do care about fairness (Fong, 2001; Corneo and Grüner, 2002; Alesina and La Ferrara, 2005; Corneo and Fong, 2008; Alesina and Giuliano, 2011; Almås et al., 2020; Fehr et al., 2021). More specifically, they underline that people tend to support greater redistribution if they believe that poverty is caused by factors beyond an individual's control, such as luck. In their broad sample of the Swiss population, Fehr et al. (2021) evaluates that half of the individuals are endowed with social preferences that are fully based on the meritocratic principle, and more than a third partly based on (consistent with the equity-efficiency trade-off). Only 15% of the individuals are found to be purely self-interested. Moreover, in finding that beliefs according to which luck rather than effort determines income[2] are strong predictors, unlike income inequality, of the national level of redistribution, Alesina, Glaeser and Sacerdote (2001) show that fair motives are of quantitative importance to explain redistributive policies. Providing a rationale for this stylised fact, in the model of Angeletos and Alesina (2005) Americans support only weak redistribution because they believe that market outcomes are fair, i.e. determined by hard work rather than luck. In their framework, as the after-tax return to effort is expected to be high, they work hard and the market outcomes are effectively fair[3]. In other words, in Alesina and Angeletos (2005) differences in redistribution are sustained because beliefs about fairness are self-fulfilling. In this article, we follow this line of explanation. However, as sustained by Luttmer and Singhal (2011), self-fulfilling beliefs alone cannot account for any persistence of welfare-state regimes. Indeed, they argue that different beliefs can be sustained over long periods only if they are embedded in culture. To that effect, we extend the approach of Alesina and Angeletos (2005) by introducing the mechanism of cultural transmission proposed by Le Garrec (2018), where culture refers to values inherited from earlier generations that translate into norms of behavior. It allows us to obtain an enriched welfare-state clustering, and thereafter to study in a meaningful way the conditions of the institutional long-term persistence.

This article is also related to the cultural economics literature that seeks to explain how be-

[2]From World Values Survey data, they highlight that 54% of Europeans versus 30% of Americans believe that luck rather than effort determines income.

[3]Note that there is no consensus on the view that market outcomes are fairer in the US than in Europe. Certainly, as reported by Alesina and Angeletos (2005), the average worked time per employee is lower in Europe than in the US. However, nothing seems to support the popular belief that American society is more mobile than European societies. Björklund and Jäntti (1997), Bratberg et al. (2017) and Helsø (2021) even show that intergenerational income mobility in Scandinavian countries would be slightly higher than in the United States. Piketty (1995) and Benabou and Tirole (2006) then explore the role of biased beliefs about social mobility to explain differences in redistribution.

liefs, tastes or preferences are formed in a society and depends on the social environment and the extent to which we see others acting in a certain way. To assess the cultural component of human behavior, recent studies have pointed out the significant and persistent differences between immigrant and native behaviors[4], or have used the different histories individuals have experienced as natural experiments[5]. Whatever the strategy employed, empirical findings support that cultural and political environment in which individuals grow up affects their preferences and beliefs concerning redistribution (Guiso et al., 2006; Alesina and Fuchs-Schündeln, 2007; Luttmer and Singhal, 2011; Alesina and Giuliano, 2011; Roth and Wohlfart, 2018). In Luttmer and Singhal (2011) for example, after controlling for individual characteristics, immigrants from countries with a preference for greater redistribution are shown to continue to give significant support to higher redistribution in their destination country. Accordingly, preferences for redistribution appear to some degree to be culturally shaped at young ages, usually referred to as *impressionable years*, and to stop changing after reaching adulthood[6]. Besides, Roth and Wohlfart (2018) show that individuals who have experienced higher levels of income inequality during their *impressionable years* support less redistribution later. It is worth noting that in Sands (2017), temporary exposure to inequality is shown to have an instantaneous (and most likely temporary) negative effect on the willingness to support redistribution. Findings of Roth and Wohlfart (2018), along with those of Luttmer and Singhal (2011), support then that it is the long-lasting exposure to inequality when young that leaves a permanent mark on the future adult's beliefs and preferences for redistribution, even if exposure to inequality is set to change. In this optic, the mechanism of cultural transmission proposed by Le Garrec (2018) specifies how, through oblique socialization, taste is shaped by the observation, imitation[7] and internalization of cultural practices[8]. More specifically, observation during childhood of high income

[4]Such as on fertility choices and women's labor supply (Fernández and Fogli, 2006), on savings (Carroll et al., 1994), on trust (Algan and Cahuc, 2010) or on preferences for redistribution (Luttmer and Singhal, 2011, Alesina and Giuliano, 2011).

[5]For example the German reunification in Alesina and Fuchs-Schündeln (2007), the Great Depression in Malmendier and Nagel (2011).

[6]Supporting this view, psychologists McCrae and Costa (1994) have shown that personality traits stop changing after age 30. See Roberts and DelVecchio (2000) and Neundorf and Smets (2017) for a discussion.

[7]In the evolutionary literature, learning from others by imitation is a cheap and efficient way to acquire locally relevant information for adaptation. Accordingly, the propensities to learn and to imitate are part of an evolved psychology shaped by natural selection (Boyd and Richerson, 1985; Boyd et al., 2011).

[8]As highlighted by the empirical findings of Dohmen et al. (2012) and discussed by Neundorf and Smets (2017), children's attitudes can also be influenced through active efforts of the parents to transmit their values, which is referred to as vertical socialization. In the theoretical litterature on cultural transmission, the socialization process is specified either by one of its channels (e.g., oblique in Le Garrec, 2018, vertical in Tabellini,

inequality, resulting from redistributive policies that are far from what would be perceived as fair, weakens concern for distributive justice[9]. The moral cost of not supporting fair taxation is reduced when observing how the previous generation has collectively failed to implement a fair institution[10]. Said differently, this mechanism states that being exposed to unfairness during youth reduces individual responsibility regarding moral duty. Typical of the literature on cultural transmission launched by Bisin and Verdier (2001), the dynamics described in Le Garrec (2018) is fundamentally backward-looking and the path leading to the different stationary states are historically dependent: to one initial condition corresponds one stationary state. In this optic, explaining the international diversity of institutions rely on understanding from where the differences in the initial conditions come from[11]. Furthermore, as part of a deterministic dynamic of history, changes in institutions are supposed to be perfectly predictable given the initial conditions. Differently, by introducing self-fulfilling beliefs within a cultural transmission framework, we shed light on the possibility that the same initial condition may lead to different stable (or quasi-stable) stationary states. In that case, it is not the differences in initial conditions that has to be understood, but the collective choices that have been made differently throughout history of different countries presenting initially similar environments. In such a framework, the institutional change appears to be hardly predictable.

The rest of the paper is organized as follows. Based on Alesina and Angeletos (2005), we present in section 2 the basic model where the multiplicity of stable equilibria results only from self-fulfilling beliefs. In section 3, we add the mechanism of cultural transmission specified in

2008), or by both of them as in the seminal work by Bisin and Verdier (2001). The latter specify that if values are sufficiently homogenous at the regional level, parents have few incentives to transmit their own values and then oblique socialization prevails.

[9]This mechanism finds support in Corneo's (2001) finding that individuals in the former West Germany showed greater concern for distributive justice than in the United States. Along the same lines, social preferences consistent with empirical distributions of wealth and income given the actual tax structure are found to give greater weight to lower-income individuals in France than in the United States (Le Grand et al., 2022).

[10]Somehow related, in the literature on crime (see Funk, 2005), it is well established that the remorse or guilt felt from breaking the social norm is weakened when observing many others are committing crimes. In the same vein, in Lindbeck et al. (1999) the individual guilt and social stigma linked to living on benefits decreases with the number of beneficiaries in the society.

[11]For example, Roland (2020) explains that the evolution towards the two current cultural systems referred to as individualism and collectivism can be traced back to different geographical conditions in antiquity. In Alesina et al. (2015), differences in family structure in the Middle Ages are at the root of current differences in labor market regulations. See Persson and Tabellini (2021) for an overview of the theoretical and empirical literature on the dynamic interactions between cultures and institutions.

Le Garrec (2018). Assimilating a welfare-state regime to a stable stationary state with perfect stable expectations, we then show that, under some conditions, there exist three configurations with three regimes. In section 4, we reexamine these results taking into account that distributive preferences are most likely heterogenous among the population. We conclude briefly in the last section.

2 The basic model

The economy is populated by a continuum of mass 1 of individuals at each generation whose actions take place according to the timeline in Figure 1. Each individual lives for two periods: childhood (*impressionable years* - developped in section 3) and adulthood. When children, they go to school to acquire knowledge and skills, making an effort maximizing their welfare. When adults, they work in order to maximize their welfare and the consumption of their household. They also vote over income redistribution.

2.1 Income and budget constraints

As underlined in the introduction, an abundant literature shows that people care about the equity of market income distribution, where factors beyond one's control such as luck characterize the level of unfairness. Accordingly, following Piketty (1995), Alesina and Angeletos (2005), Bénabou and Tirole (2006) and Le Garrec (2018), we assume that income y_{it} of an adult at date t is determined conjointly by luck and by effort such that:

$$y_{it} = A_i \left[\gamma h_{it-1} + (1-\gamma) e_{it}\right] + \varepsilon_i \tag{1}$$

where h_{it-1} denotes the person's chosen effort at school (when young), e_{it} and ε_i respectively his chosen effort at work and luck (or bad luck) when adult, $A_i \geq 0$ his talent or ability. $\gamma \in [0,1]$ is a technological parameter which characterizes the relative importance of effort at school in the income determination. As the proportion of effort chosen after the tax rate, $(1-\gamma)$ reflects the short-term sensitivity of effort with respect to the tax level. Indeed, if $\gamma = 1$, the level of effort is totally predetermined when the tax rate is chosen and the latter cannot have any effect on effort. It is assumed that $\{A_i, h_{it-1}, e_{it}\}$ are private information to agent i. ε_i is assumed unknown before the income distribution (sub-period 0 in Fig. 1) and such that $E_0[\varepsilon_i] = 0$, A_i and ε_i being independent and identically distributed (i.i.d.) across agents. In other words, when deciding his effort (at school and at work), an individual knows its return but cannot

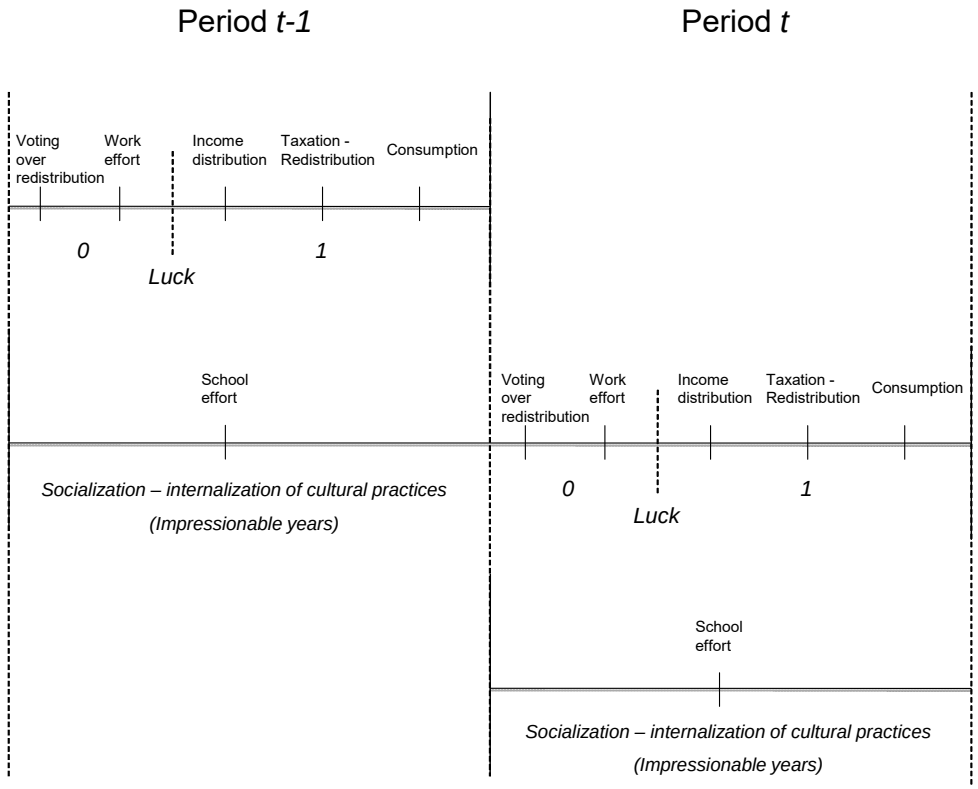

Figure 1: Timing of actions in an overlapping generations model

know if he will be lucky or not. After the income distribution, ε_i is assumed to be private information to agent i (sub-period 1 in Fig. 1).

At each period t, the government redistribute the income according to a simple fiscal scheme characterized by a flat-rate tax τ_t and a lump-sum benefit g_t provided to all adults. Assuming a balanced budget, it yields $g_t = \tau_t \bar{y}_t$, where $\bar{y}_t$ denotes the mean income at date t.

As in Boldrin and Montes (2005) and Docquier et al. (2007) children's only decision is assumed to be for education since their consumption is part of their parents' consumption. As a consequence, each adult at date t faces the following budget constraint:

$$c_{it} = y_{it}\left(1 - \tau_t\right) + \tau_t \bar{y}_t \tag{2}$$

where c_{it} denotes household consumption (one adult - one child) at date t.

2.2 Preferences and fairness

To take into account the concern for fairness that appears in redistributive attitudes, we consider an extended version of the Bolton-Ockenfels model (2000) of distributive preferences in specifying the life-cycle utility function of an individual born in period $t-1$ (adult in period

t) as:

$$U_{it-1} = u_{it-1} - \frac{\varphi}{2}\left(\tau_t^f - \tau_t\right)^2 \tag{3}$$

where u_{it-1} denotes the private life-cycle utility (from personal consumption and effort at school and at work), $\tau_t^f \in [0,1]$ the redistributive tax rate that would allow implementing the fair income distribution at date t, and $\varphi \geq 0$ the strength of the concern for fairness or inequity aversion that we assume exogenous and constant in the basic model. φ characterizes also the degree of moral universalism of preferences. In this specification, the level of redistribution perceived as fair at date t corresponds to the level of taxation optimally chosen by each adult of the same date if unfairness aversion is infinitly high: $\tau_t^f = \lim_{\varphi\to\infty} \arg\max_{\tau_t\in[0,1]} \{U_{it-1}\}$.

In the standard version of the Bolton-Ockenfels model, with no effort, the fair level of redistribution is associated to $\tau^f = 1$, i.e. characteried by equality of income. As put forward by Fehr et al. (2021), in real life, instead of pure income equality, fairness appears to be the driving principle in social preferences. Extending to (non-observable) efforts, we then follow Alesina and Angeletos (2005) in specifying the tax level perceived as fair as:

$$\tau_t^f \equiv \tau_t^{\hat{u}} = \arg\min_{\tau_t\in[0,1]} \left\{ \int_i \left(u_{it-1} - \hat{u}_{it-1}\right)^2 di \right\} \ \forall i \tag{4}$$

where $\hat{u}_{it-1}$ denotes the level of private life-cycle utility perceived as fair for an adult of type i at date t. In this setting, all the adults then share the same level of fair taxation. We will relax this assumption in section 4.

Finally, we express the private life-cycle utility as follows:

$$u_{it-1} = c_{it} - \frac{1}{2\beta_i}\left[\gamma h_{it-1}^2 + (1-\gamma)\, e_{it}^2\right] \tag{5}$$

where $\beta_i \geq 0$ characterizes a taste for effort that is assumed to be private information to agent i and i.i.d across agents. The quadratic disutility of effort is for analytical simplicity, the parameters $\frac{\gamma}{2}$ and $\frac{1-\gamma}{2}$ are a normalization. As luck is an unfair component of income, $\hat{y}_{it} = A_i\left[\gamma h_{it-1} + (1-\gamma)\, e_{it}\right]$ measures the deserved or fair income. Accordingly, the level of private life-cycle utility perceived as fair for an adult at date t is expressed as

$$\hat{u}_{it-1} = \hat{y}_{it} - \frac{1}{2\beta_i}\left[\gamma h_{it-1}^2 + (1-\gamma)\, e_{it}^2\right] \tag{6}$$

whereas the effective level, obtained with eqs. (2) and (5), is $u_{it-1} = y_{it}\left(1-\tau_t\right) + \tau_t \bar{y}_t - \frac{1}{2\beta_i}\left[\gamma h_{it-1}^2 + (1-\gamma)\, e_{it}^2\right]$.

2.3 Optimal behaviors

The optimal efforts resulting from the maximization of the expected life-cycle utility, $E_{t-1}[U_{it-1}]$ for h_{it-1} and $E_{0t}[U_{it-1}]$ for e_{it}, are as follows:

$$h_{it-1} = \beta_i A_i (1 - \tau_t^e) \tag{7}$$

$$e_{it} = \beta_i A_i (1 - \tau_t) \tag{8}$$

where τ_t^e is the anticipated (at date $t-1$) tax rate at date t. As redistribution lowers the market return to effort, it creates a disincentive to effort. In addition, as the taste for effort lowers the utility cost of effort, it enhances the effort. Considering eq. (7), the pre-tax income (1) of an adult in t can be rewritten as:

$$y_{it} = a_i [1 - \gamma\tau_t^e - (1-\gamma)\tau_t] + \varepsilon_i \tag{9}$$

where $a_i = \beta_i A_i^2$ is an index of psychological and intellectual efficiency defined thereafter as i's talent which is independant of luck. As the level of effort is reduced by redistribution, the pre-tax income is obviously also reduced. As a consequence, redistribution reduces not only the variance of the disposable income, but also the variance of the pre-tax income.

In the same optic, an adult at date t will support the level of redistribution that maximizes his utility. Assuming that the vote occurs at the beginning of the period (Fig. 1) allows the person to take into account the distortive effect of redistribution on work effort and then on income. Accordingly, considering that he can fully anticipate his future effort choice as a function of the tax rate, and evaluate the fair level of redistributive taxation as defined in eq. (4), the expected life-cycle utility (before knowing his particular luck) defined by eqs. (3) and (5) can be written, using eqs. (2), (7), (8) and (9), as:

$$\begin{aligned} E_{0t}[U_{it-1} |\ e_{it}(\tau_t)] &= [a_i(1-\tau_t) + \tau_t\bar{a}][1 - \gamma\tau_t^e - (1-\gamma)\tau_t] \\ &\quad - \frac{a_i}{2}\left[\gamma(1-\tau_t^e)^2 + (1-\gamma)(1-\tau_t)^2\right] - \frac{\varphi}{2}\left(\tau_t^f - \tau_t\right)^2 \end{aligned} \tag{10}$$

where $\bar{a}$ denotes the mean a_i. Defining the demand for redistribution of an adult at date t as the level of taxation that maximizes his utility (10) leads then to the following first order condition $(\bar{a} - a_i)[1 - \gamma\tau_t^e - 2(1-\gamma)\tau_t] - a_i(1-\gamma)\tau_t + \varphi\left(\tau_t^f - \tau_t\right) = 0$. Therefore, as long as the second order condition $a_i - 2\bar{a} - \frac{\varphi}{1-\gamma} \leq 0$ is satisfied, individual demands for redistribution at date t can be expressed as:

$$\tau_{it} = \begin{cases} \frac{(1-\gamma\tau_t^e)(\bar{a}-a_i)+\varphi\tau_t^f}{(1-\gamma)(2\bar{a}-a_i)+\varphi} & \text{if } a_i - \bar{a} \leq \frac{\varphi\tau_t^f}{1-\gamma\tau_t^e} \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

Considering the second order condition, assuming $\max_i \{a_i\} \leq 2\bar{a}$ is a sufficient condition so that preferences are single-peaked in τ_t. Individual demands for redistribution as specified in eq. (11) decrease with personal income and increase with the level of redistribution perceived as fair. From that perspective, eq. (11) is consistent with empirical surveys (Fong, 2001, Corneo and Grüner, 2002, Alesina and La Ferrara, 2005, Corneo and Fong, 2008, Alesina and Giuliano, 2011).

2.4 Policies and expectations

We now assume that, in a democracy, any policy to be implemented must be supported by a majority of adults[12]. In our model, under the sufficient condition $\max_i \{a_i\} \leq 2\bar{a}$, preferences are single-peaked in τ_t. Thus the median-voter theorem applies. Denote by $\Delta = \bar{a} - a_m$ an aggregate index of income inequality[13], where a_m denotes the median a_i, and normalize $a_m = 2$ (without loss of generality). Assuming that the distribution of (squared) talents a_i is skewed to the right yields $\Delta \geq 0$ (so that the median income is lower than the average income as observed). It follows from eq. (11) that the tax rate chosen under the majority rule can be expressed as follows:

$$\tau_{t+1} = \xi \mathcal{T}^s \left(\tau_{t+1}^e\right) + (1-\xi)\, \mathcal{T}^f \left(\tau_{t+1}^e\right) \equiv \Psi \left(\tau_{t+1}^e\right) \tag{12}$$

where $\xi = \frac{2(1-\gamma)(1+\Delta)}{2(1-\gamma)(1+\Delta)+\varphi} \in (0,1]$, $\mathcal{T}^s \left(\tau_{t+1}^e\right) = \frac{(1-\gamma\tau_t^e)\Delta}{2(1-\gamma)(1+\Delta)}$ and $\mathcal{T}^f \left(\tau_{t+1}^e\right)$ denoting respectively the preferred tax rate of the pivotal voter conditional to expectation τ_{t+1}^e if voters are either only self-interested, i.e. $\varphi = 0$, or only morally motivated, i.e. $\varphi \to +\infty$. ξ provides a measure of the proximity of the redistributive tax to the purely interested level (relatively to the purely fair level). From that perspective, it is worth noting that ξ increases as income inequality increases, $\frac{\partial \xi}{\partial \Delta} > 0$, increases as the short-term sensitivity of effort with respect to the tax level increases, $\frac{\partial \xi}{\partial (1-\gamma)} > 0$, and decreases as the concern for fairness increases, $\frac{\partial \xi}{\partial \varphi} < 0$.

The unique *selfish* equilibrium with perfect expectations, $\tau_{t+1}^e = \tau_{t+1}$, is then characterized by the following tax rate:

[12] As put forward by Corneo and Neher (2015), democracies implement to a large degree the level of redistribution demanded by the median voter.

[13] In the empirical literature investigating the link between redistribution and income inequality, the Gini coefficient or the interdecile ratios are often favored to measure income inequality.

$$\tau^s \equiv \mathcal{T}^s(\tau^s) = \frac{\Delta}{\gamma\Delta + 2(1-\gamma)(1+\Delta)} \tag{13}$$

This *selfish* tax rate exhibits the standard Meltzer-Richard (1981) effect: as income inequality rises, the median voter is poorer compared with the average and support then greater redistribution: $\frac{\partial \tau^s}{\partial \Delta} \geq 0$, where $\lim_{\Delta \to 0} \tau^s = 0$ and $\lim_{\Delta \to +\infty} \tau^s = \frac{1}{2-\gamma} \leq 1$. In addition, the tax rate decreases as voters internalize the distortion tax effect on effort and income: $\frac{\partial \tau^s}{\partial (1-\gamma)} \leq 0$, where $\lim_{\gamma \to 0} \tau^s = \frac{\Delta}{2(1+\Delta)} \leq \frac{1}{2}$ and $\lim_{\gamma \to 1} \tau^s = 1$.

Considering the purely fair tax rate $\mathcal{T}^f\left(\tau^e_{t+1}\right)$, we can show (see Appendix A) that $\lim_{\frac{\sigma^2_\varepsilon}{\sigma^2_a} \to 0^+} \mathcal{T}^f(\tau^e) = 0$ if $\tau^e < 1$ and $\lim_{\frac{\sigma^2_\varepsilon}{\sigma^2_a} \to +\infty} \mathcal{T}^f(\tau^e) = 1$. However, as the ratio $\frac{\sigma^2_\varepsilon}{\sigma^2_a}$ characterizes the perceived relative importance of luck in the income determination, these two polar cases are too trivial to be really interesting *per se*: if there is no luck, no need to redistribute on a meritocratic basis; if only luck explains the income dispersion, a total taxation may be morally justified. In addition, in these two cases, having added distributional preferences in the utility function cannot help explain why similar countries would redistribute such different proportions of their wealth. In the following, we will then focus on intermediate values of $\frac{\sigma^2_\varepsilon}{\sigma^2_a}$.

Lemma 1 *Denote by* $L_{\text{inf}} = \min\left\{2(1-\gamma)^2, \frac{1}{2}\right\}$. *If* $L_{\text{inf}} \leq \frac{\sigma^2_\varepsilon}{\sigma^2_a} < +\infty$ *there exists* $\tilde{\tau} = \max\left\{2 - \frac{1}{\gamma}, 0\right\}$ *so that* $\tau^e \geq \tilde{\tau}$ *yields* $\mathcal{T}^f(\tau^e) = 1$; $\forall \frac{1}{2} < \gamma < 1$ *and* $\tau^e < \tilde{\tau}$, *it then yields:*

- $\frac{\partial \mathcal{T}^f}{\partial \tau^e}(\tau^e) \geq 0$ *and* $0 < \lim_{\gamma \to 1} \frac{\partial \mathcal{T}^f}{\partial \tau^e}(0) \leq \frac{1}{2}$,
- $\frac{\partial \mathcal{T}^f}{\partial \frac{\sigma^2_\varepsilon}{\sigma^2_a}}(\tau^e) \geq 0$, *where* $\frac{\sigma^2_\varepsilon}{\sigma^2_a} > \frac{1}{4(2-\gamma)} \Leftrightarrow \mathcal{T}^f(\tau^e) > \tau^e$,
- $\frac{\partial \tilde{\tau}}{\partial \gamma} > 0$, $\frac{\partial \mathcal{T}^f}{\partial \gamma}(0) < 0$ *and* $\lim_{\gamma \to 1} \mathcal{T}^f(0) > 0$.

We can deduce from Lemma 1, as long as $\frac{\sigma^2_\varepsilon}{\sigma^2_a} \geq L_{\text{inf}}$ and $\frac{1}{2} < \gamma < 1$, that the higher the expected tax rate, the higher the fair-motivated tax rate[14]. Indeed, by expecting a high redistribution, individuals invest little in their human capital as its private return is low. High redistribution discourages accumulation of human capital. Therefore, the relative importance of luck in the income determination increases, as well as the moral motivation to compensate the prevalence of luck through income redistribution, until it reaches a complete taxation when $\tau^e \geq \tilde{\tau}$. This effect is analytically clear when $\gamma \to 1$. In that case, $\mathcal{T}^f(\tau^e) \to \frac{\frac{\sigma^2_\varepsilon}{\sigma^2_a}}{\frac{\sigma^2_\varepsilon}{\sigma^2_a} + (1-\tau^e)^2}$ and

[14] When $\gamma \leq \frac{1}{2}$, the tax rate perceived as fair is constant and does not depend on expectations as in Le Garrec (2018).

$\tilde{\tau} \to 1$. In addition, we can deduce from $\frac{d\tilde{\tau}}{d\gamma} > 0$ that the fair taxation tends to decrease with the relative importance of school effort compared with work effort in the income determination. Indeed, taxation is all the more efficient in reducing the income dispersion that it is internalized in the decision of private agents, i.e. that the effort is not predetermined when taxation is chosen. For example, when $\gamma = 0$ there is no predetermined effort and the obvious taxation allowing a global minimization of the fair motive as specified in eq. (4) is $\tau^f = 1 \ \forall \tau^e$.

As shown by Alesina and Angeletos (2005) and illustrated in Figure 2, with income depending on both effort and luck, and with distributive preferences reflecting the meritocratic principle of *each individual should receive what he deserves*, the cultural variability of the level of redistribution can arise as a multiplicity of stable equilibria sustained by different self-fulfilling beliefs, stable equilibria being the ones where the graph of Ψ in the (τ^e, τ) plane cuts the main diagonal from above, and unstable ones are those where it cuts it from below[15]. In that perspective, by expecting low redistribution, Americans invest in their human capital and generate conditions for low redistribution by reducing the importance of luck in the income determination. Conversely, by expecting a high redistribution, Europeans invest less in their human capital and will support a high redistribution later. Note that the existence of multiple fair-motivated tax rates, characterized by the condition $\frac{\sigma_\varepsilon^2}{\sigma_a^2} \leq \tilde{L}$, where $\tilde{L} = \frac{1}{4(2-\gamma)}$ as in Figure 2a, is neither a necessary nor a sufficient condition for the model to exhibit multiple equilibria.

Proposition 1 *Assume* $\frac{\sigma_\varepsilon^2}{\sigma_a^2} \geq L_{\inf}$, $\frac{1}{2} < \gamma < 1$, *and* $\mathcal{T}^f(\tau^e) > \mathcal{T}^s(\tau^e) \ \forall \tau^e$. *Denoting by* $\tilde{L} = \frac{1}{4(2-\gamma)}$, *the basic model exhibits two stable equilibria if:*
(1) $\frac{\sigma_\varepsilon^2}{\sigma_a^2} \leq \tilde{L}$ *and* $\varphi \geq \underline{\varphi} > 0$,
(2) $\tilde{L} < \frac{\sigma_\varepsilon^2}{\sigma_a^2} < L_{\sup}$ *and* $\varphi \in \left[\underline{\varphi}, \overline{\varphi}\right]$,
where $\frac{\partial \varphi}{\partial \frac{\sigma_\varepsilon^2}{\sigma_a^2}} \lesseqgtr 0$, $\frac{\partial \overline{\varphi}}{\partial \frac{\sigma_\varepsilon^2}{\sigma_a^2}} < 0$, $\lim\limits_{\frac{\sigma_\varepsilon^2}{\sigma_a^2} \to \tilde{L}^+} \overline{\varphi} = +\infty$ *and* $\underline{\varphi}(L_{\sup}) = \overline{\varphi}(L_{\sup})$.

For low values of $\frac{\sigma_\varepsilon^2}{\sigma_a^2}$, the lower value of φ such that the basic model exhibits two stable equilibria, denoted $\underline{\varphi}$, is associated with the threshold $\tilde{\tau}$ beyond which the perceived fair tax rate τ^f is equal to 1. Accordingly, $\underline{\varphi}$ must solve eq. (12) in which $\tau = \tau^e = \tilde{\tau}$ and is then constant (Appendix B). For higher values, it becomes slightly decreasing. As for $\overline{\varphi}$, the higher value of φ such that the basic model exhibits two stable equilibria, in order to maintain the tangency between the graph of Ψ and the main diagonal, it must decrease as the minimal distance between the perceived fair tax rate $\mathcal{T}^f(\tau)$ and the effective tax rate τ increases. As this distance tends to 0 when $\frac{\sigma_\varepsilon^2}{\sigma_a^2} \to \tilde{L}^+$, $\overline{\varphi}$ becomes infinitly high in that case (Appendix B).

[15] Local stability is characterized by the standard condition $\left|\frac{d\tau}{d\tau^e}\right| = \left|\xi \frac{\partial \mathcal{T}^s}{\partial \tau^e}(\tau) + (1-\xi) \frac{\partial \mathcal{T}^f}{\partial \tau^e}(\tau)\right| \leq 1$.

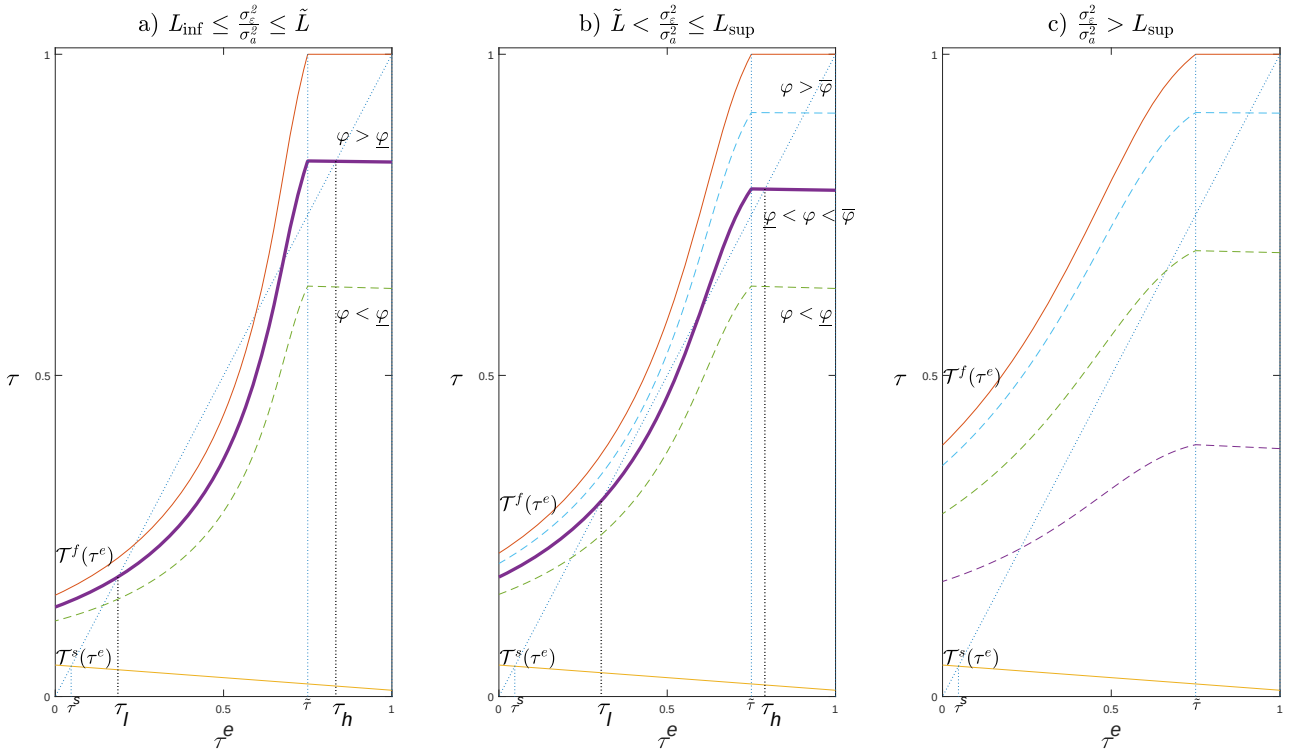

Figure 2: Redistribution and multiplicity of stable self-fullfilling beliefs ($\frac{1}{2} < \gamma < 1$)

Note finally that $\overline{\varphi}$ equals $\underline{\varphi}$ in $\frac{\sigma_\varepsilon^2}{\sigma_a^2} = L_{\sup}$ so that if $\frac{\sigma_\varepsilon^2}{\sigma_a^2} > L_{\sup}$, $\left[\underline{\varphi}, \overline{\varphi}\right] = \emptyset$ and the model cannot exhibit multiple equilibria anymore (see Fig. 2c and Fig. 7 in Appendix B).

Considering the mechanism of self-fulfilling beliefs, Luttmer and Singhal (2011) have however expressed reservations about the capacity of such beliefs to persist over long periods of time and across generations. They argue indeed that different beliefs can be sustained over long periods only if they are embedded in culture, as first conceptualized by Weber (1905) in Protestant cultures. More generally, their empirical findings, along with those of Guiso et al. (2006), Alesina and Fuchs-Schündeln (2007) and Alesina and Giuliano (2011), support that cultural and political environment in which individuals grow up affects their preferences and beliefs concerning redistribution. More specifically, they show that immigrants from high preference redistribution countries continue to support greater redistribution in their destination country. By focusing on family background as another unfair factor in the income determination, Alesina and Angeletos (2005) have extended their model to explain the history-dependency of redistributive policies. However, this extension still does not explain why immigrants from high redistribution preference countries continue to support greater redistribution. By contrast, we can observe that, if $\tau^f \geq \frac{(1-\gamma\tau)\bar{a}}{2(1-\gamma)\bar{a}}$, $\frac{\partial \tau_i}{\partial \gamma} \geq 0$. It then suggests that as long as the tax rate perceived as fair is sufficiently high, the fact that immigrants from high preference redistribution countries continue to support higher redistribution in their destination country is sustained in our

framework by a strengthened demand for social justice. In what follows, we then propose an endogenous mechanism of the formation of preferences based on internalization of the cultural practices when young in line with Le Garrec (2018).

3 Social norms and cultural transmission

3.1 Oblique socialization

To incorporate social forces into individual behavior, one privileged way is by considering the formation of agents' preferences[16]. Preferences are to some degree socially determined, so that agents internalize preferences that reflect the cultural practices of the society that they inhabit. Through oblique socialization, young individuals internalize (during their *impressionable years*), by imitation and learning, preferences that will influence their behavior when they become adults, which will explain the persistence of the cultural practices.

Assume then that the distributive preferences of an individual youth at date t are influenced by the observation of the social environment and its degree of fairness. We characterize the social environment by the social distance[17] to distributive justice:

$$\mathcal{S}_t \equiv \mathcal{S}(\tau_t) = \left[\mathcal{T}^f(\tau_t) - \tau_t\right]^2 \tag{14}$$

The higher $\mathcal{S}_t$, the more unfair the redistributive system perceived by the population. As the level of taxation τ_t results from a collective choice of the adults at date t through voting, a significant $\mathcal{S}_t$ reveals a low weight attached to the moral norm adherence and a failure in implementing fair taxation. This low weight is therefore transmitted to the young generation through observation and imitation. Having been exposed to unfairness during youth reduces the concern for fairness. Following Le Garrec (2018), we assume

$$\varphi_t = \Phi(\mathcal{S}_t),\ \Phi' \leq 0 \tag{15}$$

where φ_t is the strength of the concern for fairness developed during youth at the date t. This mechanism is closely related to that of Lindbeck et al. (1999) and Funk (2005), where the disutility of deviating from the norm is non-increasing in the fraction of deviators. However, in our setting the choice is not a binary choice between working full-time or living off benefits, as in

[16] See Postlewaite (2011) for an overview of the different approaches in the economic literature linking individual behaviors and social environment.

[17] see Akerlof (1997) for a presentation of the concept of social distance in different contexts.

Lindbeck et al. (1999), or following the law or committing a crime, as in Funk (2005). Therefore, to determine the deviation from the moral norm, the fraction of deviators is replaced by the distance between the collective choice and the norm. In addition, in our model, to characterize the socialization process, the impact on preferences of deviating from the norm applies with a delay of one generation. The moral cost of not supporting fair taxation is reduced when observing how the previous generation has collectively failed to implement a fair institution. In our framework, meritocratic fairness (linked to effort, talent and luck) is assumed to be shared by the whole population. Parents have then no incentive to transmit other values and oblique socialization naturally prevails (see Bisin and Verdier, 2001).

3.2 Multiplicity and dynamics of the welfare-state regimes

As the strength of the concern for fairness developed during youth is now endogenous and defined by eq. (15), we can write the dynamics of redistribution with social norms by adapting equation (12):

$$\tau_{t+1} = \xi_t \mathcal{T}^s\left(\tau^e_{t+1}\right) + \left(1-\xi_t\right)\mathcal{T}^{\boldsymbol{f}}\left(\tau^e_{t+1}\right) \tag{16}$$

where $\xi_t = \frac{2(1-\gamma)(1+\Delta)}{2(1-\gamma)(1+\Delta)+\Phi(\mathcal{S}_t)} \in (0,1]$ and $\Phi\left(\mathcal{S}_t\right) = \Phi\left(\left[\mathcal{T}^{\boldsymbol{f}}\left(\tau_t\right) - \tau_t\right]^2\right)$, $\tau_0 \geq 0$ given (such that the degree of moral universalism φ_0 is given). This dynamics is now obviously both history- and belief-dependent. With perfect expectations, the dynamics (16) can be expressed as:

$$\Gamma\left(\tau_{t+1}\right) \equiv \frac{\mathcal{T}^{\boldsymbol{f}}\left(\tau_{t+1}\right) - \tau_{t+1}}{\mathcal{T}^{\boldsymbol{f}}\left(\tau_{t+1}\right) - \tau^s} = \frac{\gamma\Delta + 2\left(1-\gamma\right)\left(1+\Delta\right)}{\gamma\Delta + 2\left(1-\gamma\right)\left(1+\Delta\right) + \Phi\left(\left[\mathcal{T}^{\boldsymbol{f}}\left(\tau_t\right) - \tau_t\right]^2\right)} \equiv \tilde{\xi}\left(\tau_t\right) \tag{17}$$

where $0 \leq \tilde{\xi}\left(\tau_t\right) < 1$, $\tau_0 \geq 0$ given. $\tilde{\xi}\left(\tau_t\right)$ is the transformed measure with perfect expectation of ξ_t the proximity of the redistributive tax to the purely interested level (relatively to the purely fair level): $\lim_{\tilde{\xi}\to 1}\tau = \tau^s$ and $\lim_{\tilde{\xi}\to 0}\tau = \mathcal{T}^{\boldsymbol{f}}\left(\tau\right)$. It has similar characteristics: $\frac{\partial\tilde{\xi}}{\partial\Delta} > 0$, $\frac{\partial\tilde{\xi}}{\partial(1-\gamma)} > 0$, and $\frac{\partial\tilde{\xi}}{\partial\Phi} < 0$.

3.2.1 case 1: $\left(L_{\text{inf}} \leq\right) \frac{\sigma_\varepsilon^2}{\sigma_a^2} \leq \tilde{L}$ $\left(\frac{1}{2} < \gamma < 1\right)$

To investigate the dynamics (17) and its implications in terms of stationarity defined by $\Gamma\left(\tau\right) = \tilde{\xi}\left(\tau\right)$, assume first that $2\left(1-\gamma\right)^2 \leq \frac{\sigma_\varepsilon^2}{\sigma_a^2} \leq \frac{1}{4(2-\gamma)}$, with $\frac{1}{2} < \gamma < 1$, i.e. that the graph of $\mathcal{T}^{\boldsymbol{f}}\left(\tau\right)$ cuts the main diagonal three times (Fig. 2a) or equivalently that the equation $\tau = \mathcal{T}^{\boldsymbol{f}}\left(\tau\right)$ has three positive roots $\tau_1^f \leq \tau_2^f < \tau_3^f \left(= 1\right)$.

It follows that both $\Gamma(\tau)$ and $\tilde{\xi}(\tau)$ are decreasing on $\left[0, \tau_1^f\right]$ such that $\tilde{\Psi}(\tau) = \Gamma^{-1} \circ \tilde{\xi}(\tau)$ is increasing on the same interval: $\Psi' \geq 0 \, \forall \tau \in \left[0, \tau_1^f\right]$. Consequently, on $\left[0, \tau_1^f\right]$, stable stationary states are the ones where the graph of $\tilde{\Psi}$ cuts the main diagonal from above, and unstable ones are those where it cuts it from below. On $\left[0, \tau_1^f\right]$, as eq. (17) yields $\tau_t = \tau^s \Rightarrow \tau_{t+1} \geq \tau^s$ and $\tau_t = \tau_1^f \Rightarrow \tau_{t+1} < \tau_1^f$[18], we know that there exits at least one stable stationary state $\tau_{SS} \in \left[\tau^s, \tau_1^f\right)$. As exhibited in eq. (17), the higher the concern for fairness φ, the closer the stable stationary state τ_{SS} is to the fair tax rate τ_1^f. Reciprocally, by assuming that being exposed to unfairness during youth reduces the concern for fairness, the mechanism we are exploring states also that the strength of the concern for fairness increases as the tax rate becomes fairer: $\varphi = \Phi\left(\left[\tau^f - \tau\right]^2\right)$ where $\Phi' \leq 0$. Therefore, if $\lim_{\tau \to \tau_1^f} \Phi\left(\left[\mathcal{T}^{\boldsymbol{f}}(\tau) - \tau\right]^2\right) = \Phi(0)$ is sufficiently high, there exists a stable stationary state close to the fair taxation τ_1^f. When $\lim_{\tau \to \tau_1^f} \Phi\left(\left[\mathcal{T}^{\boldsymbol{f}}(\tau) - \tau\right]^2\right) = +\infty$, it can be shown that $\tau_{SS} \to \tau_1^f$ $\left(\tau_{SS} < \tau_1^f\right)$ and $\Psi'(\tau_{SS}) \to 0$. Reasoning similarly, it is obvious that a stable stationary state is closer to the selfish tax rate the lower concern for fairness φ. If $\varphi = 0$, the unique stable stationary state is characterized by $\tau_{SS} = \tau^s$. Therefore, if $\lim_{\tau \to \tau^s} \Phi\left(\left[\mathcal{T}^{\boldsymbol{f}}(\tau) - \tau\right]^2\right)$ is sufficiently low, there exists a stable stationary state close to the selfish taxation τ^s. Accordingly, as $\Phi' \leq 0$, if both $\lim_{\tau \to \tau_1^f} \Phi\left(\left[\mathcal{T}^{\boldsymbol{f}}(\tau) - \tau\right]^2\right)$ is sufficiently high and $\lim_{\tau \to \tau^s} \Phi\left(\left[\mathcal{T}^{\boldsymbol{f}}(\tau) - \tau\right]^2\right)$ is sufficiently low, two stable stationary states coexist on $\left[\tau^s, \tau_1^f\right)$ characterized by tax rates τ_L and τ_I, $\tau^s \leq \tau_L < \tau_I < \tau_1^f$. Note that, as the graph of $\tilde{\Psi}$ can cross the diagonal an odd number of times which may be greater than 3, the number of stable stationary states may be greater than 2 (we do not consider here the non-generic case of uncountable many crossings). Restricting our attention to the case with only two stable stationary states on $\left[\tau^s, \tau_1^f\right]$ requires that we assume (regularity conditions), in addition, that Φ is of class C^2 and that the equation $\tilde{\Psi}'(\tau) = 1$ has only two roots on $\left[0, \tau_1^f\right]$.

Similarly, as both $\Gamma(\tau)$ and $\tilde{\xi}(\tau)$ are decreasing on $\left[\tilde{\tau}, \tau_3^f\,(=1)\right]$, $\tilde{\Psi}\left(=\Gamma^{-1} \circ \tilde{\xi}\right)$ is increasing on the same interval: $\tilde{\Psi}' \geq 0 \; \forall \tau \in \left[\tilde{\tau}, \tau_3^f\right]$. Following the same reasoning as above, if $\lim_{\tau \to \tau_3^f} \Phi\left(\left[\mathcal{T}^{\boldsymbol{f}}(\tau) - \tau\right]^2\right) = \Phi(0)$ is sufficiently high, there exists a stable stationary state close to the fair taxation τ_3^f, with tax rate τ_H, where $\tau_H \to \tau_3^f$ $\left(\tau_H < \tau_3^f\right)$ and $\tilde{\Psi}'(\tau_H) \to 0$ when $\Phi(0) \to +\infty$. By definition, another stable stationary state close to the selfish taxation cannot be added as $\tau^s < \tilde{\tau}$. This is not the case for the second root τ_2^f of the equation $\tau = \mathcal{T}^{\boldsymbol{f}}(\tau)$ if $\lim_{\tau \to \tau_2^f} \Phi\left(\left[\mathcal{T}^{\boldsymbol{f}}(\tau) - \tau\right]^2\right) = \Phi(0)$ is sufficiently high[19]. However, as can be seen on Fig. 2a, this stationary state is associated with an unstable belief and so is excluded of our notion of stable

[18] More exactly $\tau_{t+1} < \mathcal{T}^{\boldsymbol{f}}(\tau_{t+1})$ where we assume $\mathcal{T}^{\boldsymbol{f}}(\tau_{t+1}) \leq \tau_1^f$ to stay on the interval $\left[0, \tau_1^f\right]$.

[19] In that case $\tilde{\Psi}$ is decreasing and its graph crosses the main diagonal with a slope less than unity in absolute

fiscal regime. About stable beliefs, one can also observe on Fig. 2a that if the high-tax stable equilibrium exists, so does the low one, i.e. that φ is necessarily greater that $\underline{\varphi}$ (Proposition 1). We deduce that $\Phi\left(\left[\mathcal{T}^{\boldsymbol{f}}(\tau_H)-\tau_H\right]^2\right)\geq\underline{\varphi}$ and that two stable self-fulfilling beliefs associated with history $\tau_t=\tau_H$ exist such that: $\tau_t=\tau_H\Rightarrow\tau_{t+1}=\tau_{t+1}^e=\left\{\hat{\tau}\in\left(\tau_1^f,\tau_H\right),\tau_H\right\}$. To illustrate this property as simply as possible, the dynamics (16) can be expressed when $\Phi(0)\to+\infty$ as: $\lim_{\tau_t\to\tau_3^f,\Phi(0)\to+\infty}\tau_{t+1}=\mathcal{T}^{\boldsymbol{f}}\left(\tau_{t+1}^e\right)=\tau_{t+1}^e=\left\{\tau_1^f,\tau_3^f\right\}$. As a stable self-fulfilling belief can potentially destabilize the high-redistributive stationary state characterized by the tax rate τ_H, the latter is only a weakly-stable regime, stable in local-belief, i.e. only locally determinate but not globally. Considering the intermediate-redistribution stationary state characterized by the tax rate τ_I, we can show (see Appendix C) that its distance to the purely fair redistributive tax rate $\mathcal{T}^{\boldsymbol{f}}(\tau_I)$ is lower than in the high-redistributive stationary state. Therefore, $\Phi\left(\left[\mathcal{T}^{\boldsymbol{f}}(\tau_I)-\tau_I\right]^2\right)\geq\Phi\left(\left[\mathcal{T}^{\boldsymbol{f}}(\tau_{SD})-\tau_{SD}\right]^2\right)\geq\underline{\varphi}$ so that two stable self-fulfilling beliefs associated with history $\tau_t=\tau_I$ exist: $\tau_t=\tau_I\Rightarrow\tau_{t+1}=\tau_{t+1}^e=\left\{\tau_I,\check{\tau}\in\left(\tau_H,\tau_3^f\right)\right\}$. The intermediate-redistribution stationary state is also a weakly-stable regime.

To summarize our findings, verifying some regularity conditions (to minimize the number of stationary states) and assimilating a welfare-state regime to a stable stationary state with perfect stable expectations, we formulate the following proposition:

Proposition 2 *Assume that* $L_{\inf}\leq\frac{\sigma_\varepsilon^2}{\sigma_a^2}\leq\tilde{L}$, *with* $\frac{1}{2}<\gamma<1$. *If* $\varphi=\Phi(\mathcal{S})$ *is sufficiently high when* τ *approaches* $\mathcal{T}^{\boldsymbol{f}}(\tau)$ *and sufficiently low when* τ *approaches* τ^s, *there exist three welfare-state regimes characterized by tax rates* $\tau_L<\tau_I<\tau_H$, *where the intermediate- and high-redistribution regimes are only stable in local-belief, or weakly stable, with* $\frac{\partial\tau_L}{\partial\Delta}>0$, $\frac{\partial\tau_I}{\partial\Delta}>0$, $\frac{\partial\tau_{SD}}{\partial\Delta}>0$, $\frac{\partial\tau_L}{\partial\frac{\sigma_\varepsilon^2}{\sigma_a^2}}<0$, $\frac{\partial\tau_I}{\partial\frac{\sigma_\varepsilon^2}{\sigma_a^2}}>0$ *and* $\frac{\partial\tau_H}{\partial\frac{\sigma_\varepsilon^2}{\sigma_a^2}}=0$.

To illustrate this proposition, take the following function $\Phi(\mathcal{S})=\frac{\alpha}{\beta+\mathcal{S}^\theta}$, where α, β and θ are three strictly positive parameters. As is obvious, the concern for fairness $\varphi=\frac{\alpha}{\beta+\left[\tau^f-\tau\right]^{2\theta}}$ is high when τ approaches τ^f if $\frac{\alpha}{\beta}$ is high, and low when τ approaches $\tau^s<\tau^f$ if α is low. Compared with Proposition 2, the condition $\Phi(\mathcal{S})$ is sufficiently high when τ approaches τ^f can be redefined as $\beta\leq\hat{\beta}$ (and then $\frac{\alpha}{\beta}$ sufficiently high where $\lim_{\beta\to0,\beta>0}\Phi(0)=+\infty$) and the condition $\Phi(\mathcal{S})$ is sufficiently low when τ approaches τ^s as $\alpha\leq\hat{\alpha}$. As illustrated in Fig. 3a, if both $\alpha\leq\hat{\alpha}$ and $\beta\leq\hat{\beta}$ there exist three stable stationary state with perfect stable expectations characterized by tax rates $\tau_L<\tau_I<\tau_H$. Note that $\tau_I\in[\tau_1,\tau_2]$ and $\tau_H\in[\tau_3,\tau_4]$, two intervals in which the social distance to distributive justice is lower than the threshold $\hat{\mathcal{S}}$. In these two intervals, the strength of the concern for fairness is then high enough such that two stable

value.

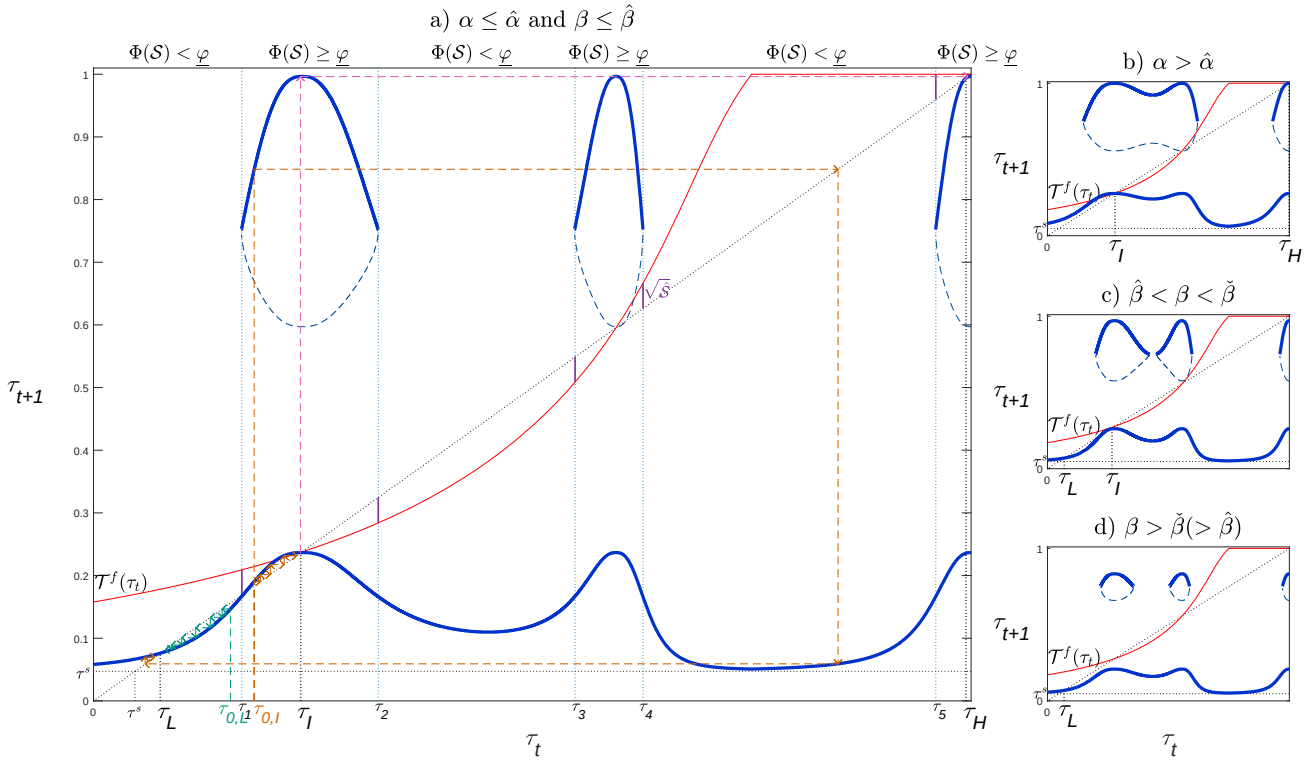

Figure 3: The quasi-stability of the three-regime clustering when $(L_{\inf} \leq) \frac{\sigma_\varepsilon^2}{\sigma_a^2} \leq \tilde{L}$ $\left(\frac{1}{2} < \gamma < 1\right.$; dashed line is for unstable beliefs$)$

expectations exist $\left(\Phi\left(\mathcal{S}_x\right) \geq \underline{\varphi},\, x = I, H\right)$. The intermediate- and high-redistribution regimes are only locally determinate, stable only in local-belief. By contrast, in τ_L, the social distance to distributive justice is higher than the threshold $\hat{\mathcal{S}}$, i.e. $\Phi\left(\mathcal{S}_L\right) < \underline{\varphi}$. The low-redistribution regime is globally determinate, i.e. fully stable. Compared to that configuration, if $\alpha > \hat{\alpha}$ the strength of the concern for fairness is always too strong for the stable stationary state τ_L to exist (Fig. 3b). If $\beta > \hat{\beta}$, two configurations may exist. First, if β is only slightly higher than $\hat{\beta}$, $\left(\beta - \hat{\beta}\right) \in \left(0, \check{\beta} - \hat{\beta}\right]$, only the high-redistribution welfare-state regime τ_H vanishes (Fig. 3c and Appendix C). Second, if β is significantly higher than $\hat{\beta}$, $\left(\beta - \hat{\beta}\right) > \left(\check{\beta} - \hat{\beta}\right)$, both intermediate- and high-redistribution regimes vanish so that only the low-redistribution regime persists (Fig. 3d). When $\beta \to +\infty$, $\Phi\left(\mathcal{S}\right) \to 0$ and we are back to a traditional model with only selfish preferences so that the only welfare-state regime is characterized by $\tau_L \to \tau^s$.

As illustrated in Figure 3a, if at date $t = 0$ the tax rate is $\tau_{0,L}$, people are socialized in an environment where practices and institutions are too far from what is perceived as fair. Thereafter, the level of taxation decreases at date $t = 1$, $\tau_{1,L} < \tau_{0,L}$, reducing the relative importance of luck in the income determination. Accordingly, the level of taxation perceived as fair decreases, $\mathcal{T}^{\boldsymbol{f}}\left(\tau_{1,L}\right) < \mathcal{T}^{\boldsymbol{f}}\left(\tau_{0,L}\right)$, but only slightly so that the perceived unfairness of the institutions increases, $\mathcal{S}_{1,L} = \left[\mathcal{T}^{\boldsymbol{f}}\left(\tau_{1,L}\right) - \tau_{1,L}\right]^2 > \left[\mathcal{T}^{\boldsymbol{f}}\left(\tau_{0,L}\right) - \tau_{0,L}\right]^2 = \mathcal{S}_{0,L}$. The

generation that is young at date $t = 1$ is socialized in an environment that is further away from the fair institution than was the previous generation. Hence, by being exposed to more unfairness, their concern for fairness decreases and they will support an institution that will be even further away from fairness at date $t = 2$. This cultural transmission process ends with the implementation of the low-redistribution institution characterized by the tax rate τ_L. The redistributive institution and the concern for fairness co-evolve and are self-reinforcing such that $\lim_{t\to+\infty} \tau_{t,L} = \tau_L$ and $\lim_{t\to+\infty} \Phi(\mathcal{S}_{t,L}) = \varphi_L$ and the process is only history dependent.

By contrast, if the initial taxation is in $\tau_{0,I}$ sufficiently close to what is perceived as fair such that $\mathcal{S}_{0,I} = \left[\mathcal{T}^{\boldsymbol{f}}(\tau_{0,I}) - \tau_{0,I}\right]^2 < \hat{\mathcal{S}}$ and then $\Phi(\mathcal{S}_{0,I}) > \underline{\varphi}$, two stable self-fulfilling beliefs are associated. If the selected belief is (and always is) in the neighborhood of the previous equilibrium, the process encountered above is reversed and the concern for fairness, the level of redistribution as well as the one perceived as fair increase with time to stabilize towards the intermediate levels $\lim_{t\to+\infty} \tau_{t,I} = \tau_I$ and $\lim_{t\to+\infty} \Phi(\mathcal{S}_{t,I}) = \varphi_I > \varphi_L$. Otherwise, the level of taxation may jump to the higher level of taxation $\tau_{1,I} = \tau_h$. In such a case, luck in the income determination becomes so dominant that even if the level of taxation is high, it is too low compared to the level perceived as fair. The generation that is young at date $t = 1$ is then socialized in an environment perceived paradoxically as very unfair and will implement the following period an institution whose level of taxation will be close to the low one. The process will end by implementing the low-redistribution institution: $\lim_{t\to+\infty} \tau_{t,I} = \tau_L$ and $\lim_{t\to+\infty} \Phi(\mathcal{S}_{t,I}) = \varphi_L$. This example illustrates that the dynamics described by eq. (17) is both history- and belief-dependent if the conditions of Proposition 2 are satisfied.

As stressed in Proposition 2, only the low-redistribution is truly stable, the other two being associated with destabilizing self-fulfilling beliefs. More specifically, starting from the intermediate state, the tax rate can jump up from τ_I to a level belonging to $\left(\tau_H, \tau_3^f\right)$ because $\Phi(\mathcal{S}_I) > \Phi(\mathcal{S}_H)$, and then converges towards τ_H. On the one hand, any institution sufficiently close to the intermediate institution can converge towards the high-redistribution regime. However, it does not mean that in a finite time the clustering of institutions would end with only two regimes, the low- and the high-redistribution ones. Indeed, if the intermediate regime can converge towards the high-redistribution one, the reverse is also possible. The clustering in three regimes appears stable in the long term in this configuration, even if the sets of countries composing the intermediate and the high-redistribution regimes can change. This characteristic underlines the qualitative similarity of the two latter regimes, both of them implementing an income redistribution being closer to fair level than in the liberal model because citizens within exhibit a stronger aversion for unfairness. It also illustrates the possibility that the welfare state

of similar countries starting from the same initial condition such that $\tau_{0,I}$ can converge towards one of the three regimes according to collective choices that are not exclusively dependent on history and therefore difficult to predict.

Focusing on fiscal features, it brings some rationale why most countinental Europeans countries appear currently in the social-democratic welfare-state regime (Péligry and Ragot, 2022) whereas they belonged to the intermediate regime 30 years ago (Esping-Andersen, 1990), and reciprocally why Norway is now clustered in the intermediate welfare-state regime. These results are intrinsically linked to the case in which the graph of $\mathcal{T}^{\boldsymbol{f}}(\tau)$ cuts the main diagonal three times that characterizes moderate unfairness in the income determination. To that extent, whether they extend to greater unfairness remains to be investigated, i.e. in case the graph of $\mathcal{T}^{\boldsymbol{f}}(\tau)$ cuts the main diagonal only once. And if not, what would remain of the multiplicity of welfare-state regimes.

3.2.2 case 2: $\tilde{L} < \frac{\sigma_\varepsilon^2}{\sigma_a^2} (< L_{\sup})$ $\left(\frac{1}{2} < \gamma < 1\right)$

As put forward in Lemma 1, if $\frac{\sigma_\varepsilon^2}{\sigma_a^2} > \tilde{L}$, then the graph of $\mathcal{T}^{\boldsymbol{f}}(\tau)$ cuts the main diagonal only once in $\tau = 1$. It drastically changes the configuration compared to the previous section as a too strong aversion for inequality ($\varphi > \overline{\varphi}$) does not yield multiple self-fulfilling beliefs anymore (Proposition 1). However, for multiple stationary state to exist, we need to continue to assume that $\varphi = \Phi(\mathcal{S})$ is sufficiently high when τ approaches $\mathcal{T}^{\boldsymbol{f}}(\tau)$ (i.e. when $\mathcal{S}$ tends towards 0) and sufficiently low when τ approaches τ^s. Hence, we ensure that the low- and high-redistribution regimes exist.

How about the existence of the intermediate regime? To study this question, define $\hat{\tau}$ the level of taxation that minimizes the observed unfairness of the redistributive institution on the interval $[0, \tilde{\tau}]$: $\hat{\tau} = \underset{\tau \in [0,\tilde{\tau}]}{\arg\min} \mathcal{S}$. In the previous case, $\hat{\tau} = \left\{\tau_1^f, \tau_2^f\right\}$ and $\mathcal{S}(\hat{\tau}) = 0$. In the present case, if an interior solution exists, i.e. if $\frac{\sigma_\varepsilon^2}{\sigma_a^2}$ is sufficiently close to $\tilde{L}$, it corresponds to $\hat{\tau} = \frac{\partial \mathcal{T}^{\boldsymbol{f}}}{\partial \tau}^{-1}(1)$, where $\mathcal{S}(\hat{\tau}) \left(= \left[\mathcal{T}^{\boldsymbol{f}}(\hat{\tau}) - \hat{\tau}\right]^2\right) > 0$, $\frac{\partial \mathcal{S}(\hat{\tau})}{\partial \frac{\sigma_\varepsilon^2}{\sigma_a^2}} > 0$ and $\lim_{\frac{\sigma_\varepsilon^2}{\sigma_a^2} \to \tilde{L}^+} \mathcal{S}(\hat{\tau}) = 0$. As $\lim_{\frac{\sigma_\varepsilon^2}{\sigma_a^2} \to \tilde{L}^+} \overline{\varphi} = +\infty$ (Proposition 1), the following Lemma holds:

Lemma 2 *Assume* $\frac{\sigma_\varepsilon^2}{\sigma_a^2} > \tilde{L}$, *with* $\frac{1}{2} < \gamma < 1$.

1. *If* $\lim_{\frac{\sigma_\varepsilon^2}{\sigma_a^2} \to \tilde{L}^+} \Phi(\mathcal{S}(\hat{\tau})) = \Phi(0)$ *is sufficiently high (yet bounded) and* $\lim_{\frac{\sigma_\varepsilon^2}{\sigma_a^2} \to \tilde{L}^+} |\Phi'(\mathcal{S}(\hat{\tau}))| < +\infty$ *there exists* $\hat{L} > \tilde{L}$ *such that* $\frac{\sigma_\varepsilon^2}{\sigma_a^2} \leqq \tilde{L}$ *yields* $\Phi(\mathcal{S}(\hat{\tau})) \leqq \overline{\varphi}\left(\frac{\sigma_\varepsilon^2}{\sigma_a^2}\right)$.

2. *If in addition* $\lim_{\frac{\sigma_\varepsilon^2}{\sigma_a^2}\to L_{\sup}^-} \Phi(\mathcal{S}(\hat{\tau})) < \underline{\varphi}$, *there exist* $\underline{L}$ *and* $\overline{L}$ *such that* $\tilde{L} < \hat{L} < \underline{L} < \overline{L} < L_{\sup}$, *where* $\underline{L}$ *is a root of equation* $\Phi(\mathcal{S}(\hat{\tau})) = \overline{\varphi}\left(\frac{\sigma_\varepsilon^2}{\sigma_a^2}\right)$ *and* $\overline{L}$ *of equation* $\Phi(\mathcal{S}(\hat{\tau})) = \underline{\varphi}\left(\frac{\sigma_\varepsilon^2}{\sigma_a^2}\right)$ (Fig. 4b).

For the function $\Phi(\mathcal{S}) = \frac{\alpha}{\beta+\mathcal{S}^\theta}$, the conditions for the existence of the threshold $\hat{L}$ are $\frac{\alpha}{\beta}$ sufficiently high and $\theta \geq 1$ (Fig. 4b). The first part of Lemma 2 stresses that the results of Proposition 2 can be extended on the interval $\left(\tilde{L}, \hat{L}\right)$, i.e. they do not strictly rely on the graph of $\mathcal{T}^{\boldsymbol{f}}(\tau)$ crossing the main diagonal three times (Fig. 4a).

As long as $\frac{\sigma_\varepsilon^2}{\sigma_a^2} \leq \hat{L}$, our results so far stress that the high-redistribution regime, if it exists, is weakly stable. However, in $\frac{\sigma_\varepsilon^2}{\sigma_a^2} = \hat{L}$, one may observe that $\Phi(0) \approx \overline{\varphi}\left(\hat{L}\right)$ (especially when $\Phi(0)$ is high enough). Therefore, if $\frac{\sigma_\varepsilon^2}{\sigma_a^2} > \hat{L}$, a fundamental change occurs as the high-redistribution regime becomes fully stable. Indeed, $\varphi = \Phi(\mathcal{S})$ sufficiently high when τ approaches $\mathcal{T}^{\boldsymbol{f}}(\tau)$ supposes in that case that $\lim_{\mathcal{S}\to 0}\Phi(\mathcal{S}) > \overline{\varphi}$. $\mathcal{S}$ sufficiently small such that $\Phi(\mathcal{S}(\tau_{SD})) > \overline{\varphi}$ then yields that in $\tau_t = \tau_H$ only the self-fulfilling belief $\tau_{t+1}^e = \tau_H$ is consistent with eq. (16), and the high-redistribution regime is globally determinate.

Consider now the successive configurations underlined by Lemma 2 when $\frac{\sigma_\varepsilon^2}{\sigma_a^2} \in \left(\hat{L}, L_{\sup}\right)$. First, if $\frac{\sigma_\varepsilon^2}{\sigma_a^2} \in \left(\hat{L}, \underline{L}\right)$, $\Phi(\mathcal{S}(\hat{\tau})) > \overline{\varphi}$ yields that in $\tau_t = \hat{\tau}$ only the belief in a high redistributive state at date $t+1$ is consistent with eq. (16). Therefore, a stationary state cannot exist in $\tau = \hat{\tau}$ (Fig. 4c). As far as τ deviates from $\hat{\tau}$, in one way or the other, the perceived unfairness of the redistributive institution $\mathcal{S}$ is increasing and the strength of the concern for fairness $\Phi(\mathcal{S})$ is decreasing. When the latter reaches $\Phi(\mathcal{S}) = \overline{\varphi}$, another self-fulfilling belief appears which is characterized by the tangence between the main diagonal and the curve of Ψ in the (τ^e, τ) plane (see Fig. 9a in Appendix D). In the (τ_t, τ_{t+1}) plane, the slope of the curve of $\tilde{\Psi}$ goes to infinity in that particular point, and we can assert that around $\hat{\tau}$ there cannot exist a stable stationary state because the curve crosses the main diagonal whith a slope higher that unity in absolute value. In Fig. 4c, it is in addition associated with an unstable belief, but it is not necessary the case.

If $\frac{\sigma_\varepsilon^2}{\sigma_a^2} \in \left[\underline{L}, \overline{L}\right]$, it yields that $\Phi(\mathcal{S}(\hat{\tau})) \in \left[\underline{\varphi}, \overline{\varphi}\right]$. However, it is not a sufficient condition for the intermediate welfare regime to exist or to be stable. Take for example $\frac{\sigma_\varepsilon^2}{\sigma_a^2} = \underline{L}$ in which $\Phi(\mathcal{S}(\hat{\tau})) = \overline{\varphi}(\underline{L})$. In that case, in $\tau_t = \hat{\tau}$ the two stable beliefs exist, both of them being higher than $\hat{\tau}$: $\tau_t = \hat{\tau} \Rightarrow \tau_{t+1} = \tau_{t+1}^e = \{\tau_l, \tau_h\}$, where $\tau_h > \tau_l > \hat{\tau}$ (see Appendix D). As the curve of $\tilde{\Psi}$ is decreasing on $[\hat{\tau}, \tilde{\tau}]$, with $\Phi(\mathcal{S})$ being small when τ_t deviates from $\hat{\tau}$, it crosses the main diagonal on that interval, which defines a stationary state. However, in this point, the slope of $\tilde{\Psi}$ is higher than unity in absolute value and the stationary state is unstable (Fig. 5b). Indeed,

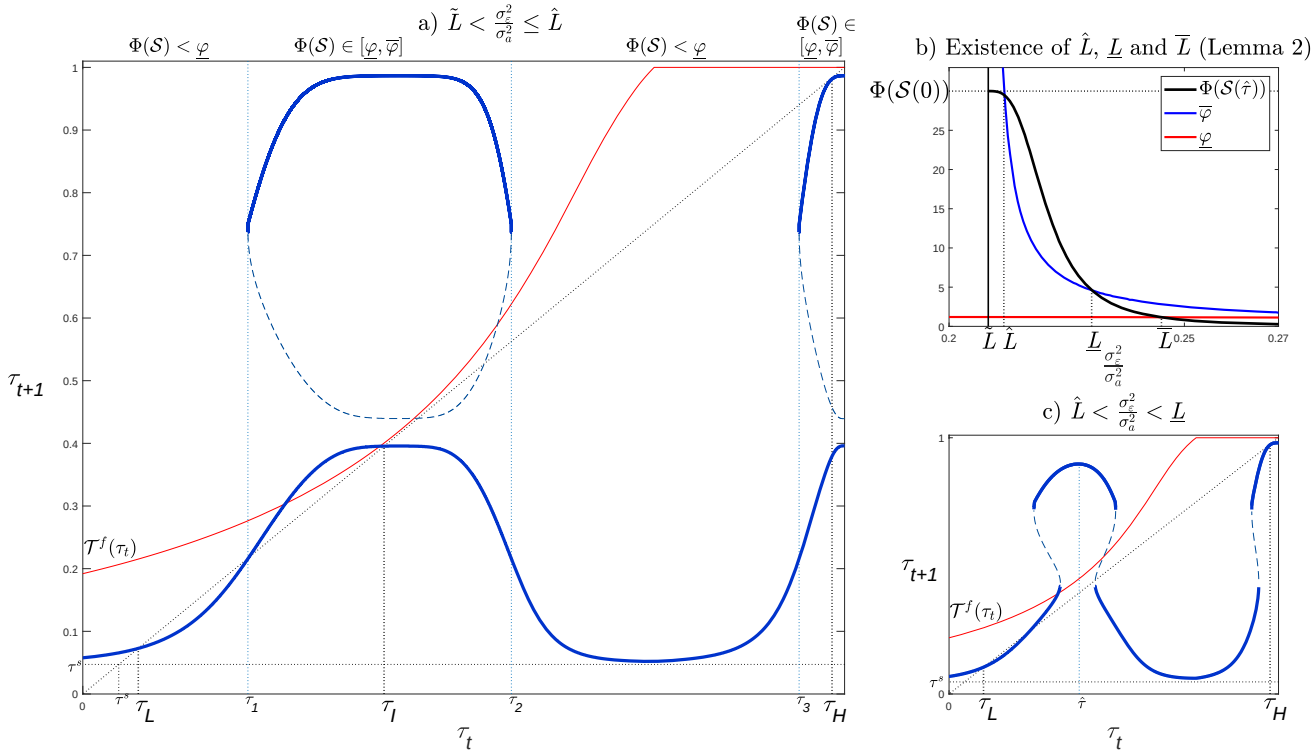

Figure 4: The quasi-stability of the three-regime clustering extended on $\frac{\sigma_\varepsilon^2}{\sigma_a^2} \in \left(\tilde{L}, \underline{L}\right)$ when Lemma 2 applies $\left(\frac{1}{2} < \gamma < 1\right.$; dashed line is for unstable beliefs$\left.\right)$

as explained above, when $\Phi(\mathcal{S}) = \overline{\varphi}$, in the (τ_t, τ_{t+1}) plane, the slope of the curve of $\tilde{\Psi}$ goes to infinity in that particular point. Take now $\frac{\sigma_\varepsilon^2}{\sigma_a^2} = \overline{L}$ in which $\Phi(\mathcal{S}(\hat{\tau})) = \underline{\varphi}\left(\overline{L}\right)$. If $\underline{\varphi}\left(\overline{L}\right)$ is too close to $\overline{\varphi}\left(\overline{L}\right)$, i.e. that $\overline{L}$ is close to $L_{\sup}$, it follows that in $\tau_t = \hat{\tau}$, both tax rates τ_l and τ_h are still higher than $\hat{\tau}$ (see Appendix D) and the stability of the stationary state cannot be guaranteed for the reasons stated above. By contrast, if $\underline{\varphi}\left(\overline{L}\right)$ is sufficiently lower than $\overline{\varphi}\left(\overline{L}\right)$, i.e. that $\overline{L}$ is far enough of $L_{\sup}$, it follows that in $\tau_t = \hat{\tau}$, $\tau_{t+1} = \tau_{t+1}^e = \{\tau_l, \tau_h\}$, where $\tau_l < \hat{\tau} < \tau_h$. Accordingly, if τ_l is sufficiently smaller than $\hat{\tau}$ in the (τ_t, τ_{t+1}) plane, the graph of $\tilde{\Psi}$ does not cross the main diagonal around $\hat{\tau}$ and there there cannot exist an intermediate stationary state (Fig. 5c). In that case, by continuity, the following Proposition holds:

Proposition 3 *Assume* $\frac{1}{2} < \gamma < 1$, $\varphi = \Phi(\mathcal{S})$ *sufficiently high when* τ *approaches* $\mathcal{T}^f(\tau)$ *and sufficiently low when* τ *approaches* τ^s. $\overline{L}$ *being sufficiently far from* $L_{\sup}$, *there exists a non-empty interval* $\left[\underline{\underline{L}}, \overline{\underline{L}}\right] \subset \left(\underline{L}, \overline{L}\right)$ *such that* $\frac{\sigma_\varepsilon^2}{\sigma_a^2} \in \left[\underline{\underline{L}}, \overline{\underline{L}}\right]$ *yields that three welfare-state regimes characterized by tax rates* $\tau_L < \tau_I < \tau_H$ *exist, where the intermediate-redistribution regime is only stable in local-belief, or weakly stable.*

In this configuration, the low- and high-redistribution regimes are globally determined, then fully stable (Fig. 5a). Only the intermediate regime is associated with destabilizing self-fulfilling

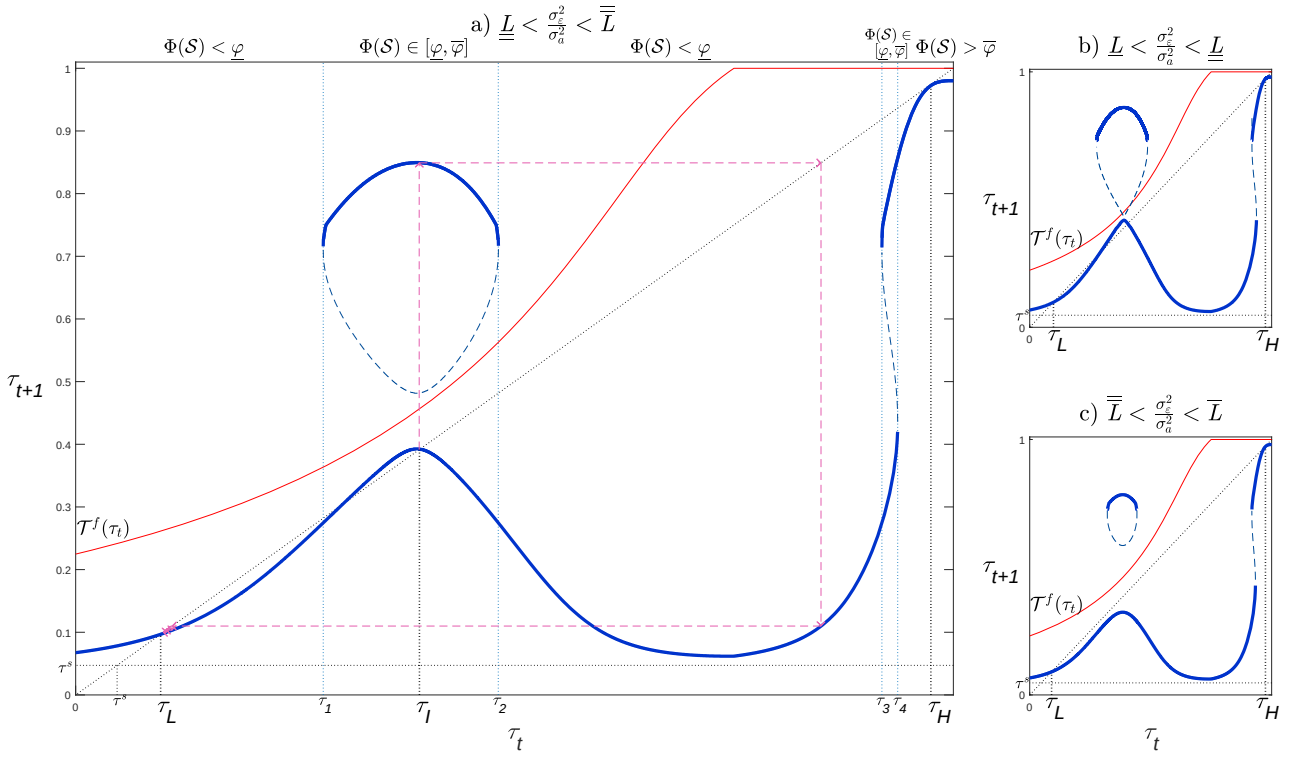

Figure 5: The long-term vanishing of the intermediate welfare-state regime when $\frac{\sigma_\varepsilon^2}{\sigma_a^2} \in \left[\underline{\underline{L}}, \overline{\overline{L}}\right]$ $\left(\frac{1}{2} < \gamma < 1\right.$; dashed line is for unstable beliefs$\left.\right)$

beliefs, i.e. only locally determined. It has important consequences in terms of long-term stability of the three-regimes clustering and the existence of the intermediate welfare-state regime. Indeed, in a finite time, we can figure out that the intermediate regime could be destabilized with self-fulfilling beliefs sustaining greater redistribution. In that case, the period after the jump, redistribution becomes effectively greater. However, contrary to the configuration previously studied, the jump up is too small to initiate a process of convergence towards the high-redistribution regime. In this one-period after the jump, the leading phenomenum is that the implemented redistributive institution is too small compared to the one that is perceived as fair. As a consequence, the generation that is young in this period is socialized in an environement whose practices and institutions are far from reflecting fairness, and internalization of the observed norm *you should behave according to your own interest* reduces the strength of the concern for fairness. A process of convergence towards the low-redistribution regime is initiated. This configuration predicts then that, in a finite time, countries with an intermediate welfare regime will end with a low-redistribution one, and the final clustering will encounter only two regimes, the low- and the high-redistribution, which cannot be destibilized by other beliefs. It underlines that in the intermediate regime, individuals are still sufficiently close to a fair institution for believing that higher redistribution is possible, but no longer close enough

to converge towards it and a fairer institution. If we favor this configuration to interpret the passage of some continental European countries from an intermediate regime to a more generous regime, the latter could have been wrongly assimilated to the social-democratic regime whereas it only characterizes a transitory situation which on the contrary leads to less social protection in the long term.

Consider finally the case where $\frac{\sigma_\varepsilon^2}{\sigma_a^2} \in \left(\overline{L}, L_{\sup}\right)$. As $\Phi\left(\mathcal{S}\left(\hat{\tau}\right)\right) < \underline{\varphi}$, it follows that in $\tau_t = \hat{\tau}$ only the belief in a low redistributive state characterized by τ_l at date $t+1$ is consistent with eq. (16). Again, two configurations arise depending on the relative distance of $\overline{L}$ from $L_{\sup}$. As explained above, if this distance is high enough in $\frac{\sigma_\varepsilon^2}{\sigma_a^2} \to \overline{L}^+$ such that τ_l is sufficiently smaller than $\hat{\tau}$ in the (τ_t, τ_{t+1}) plane, the graph of $\tilde{\Psi}$ does not cross the main diagonal around $\hat{\tau}$ and there cannot exist any intermediate stationary state in $\frac{\sigma_\varepsilon^2}{\sigma_a^2} \in \left(\overline{L}, L_{\sup}\right)$ (Fig. 6b). By contrast, if the distance between $\overline{L}$ and $L_{\sup}$ is low enough such that τ_l is higher (or only slightly smaller) than $\hat{\tau}$, the graph of $\tilde{\Psi}$ crosses the main diagonal around $\hat{\tau}$ and an intermediate stable stationary state can exist in $\frac{\sigma_\varepsilon^2}{\sigma_a^2} \in \left(\overline{L}, L_{\sup}\right)$ (Fig. 6a). In contrast to the previous cases, the latter is truly stable as no other self-fulfilling belief exists to destabilize it. The following Proposition then holds:

Proposition 4 *Assume that* $\overline{L} < \frac{\sigma_\varepsilon^2}{\sigma_a^2} < L_{\sup}$, $L_{\sup} - \overline{L}$ *being small enough, with* $\frac{1}{2} < \gamma < 1$. *If* $\varphi = \Phi\left(\mathcal{S}\right)$ *is sufficiently high when* τ *approaches* $\mathcal{T}^{\boldsymbol{f}}\left(\tau\right)$ *and sufficiently low when* τ *approaches* τ^s, *there exist three welfare-state regimes characterized by tax rates* $\tau_L < \tau_I < \tau_{SD}$.

In this configuration, the clustering of the welfare states in three regimes is supposed to persist over time. More specifically, in the intermediate regime, compared to the previous configuration, individuals are no more sufficiently close to a fair institution for believing that higher redistribution is an option. The strength of the concern for fairness is too low. It creates determinacy, as in the low-redistribution regime. It is the reverse of the high-redistribution regime in which it is the proximity with the fair institution that prevent lower redistribution to be a credible alternative. The strength of the concern for fairness is too high for that. Given the rise in inequalities observed in recent years, it is possible that, after a period of reallocation of the countries with generous redistributive institutions between the intermediate and high-redistribution regimes, a truly stable clustering of the welfare states in three regimes will appear. Note however that conditions for this configuration to exist are restrictive since the interval $\left(\overline{L}, L_{\sup}\right)$ in which $\frac{\sigma_\varepsilon^2}{\sigma_a^2}$ has to belong needs to be small.

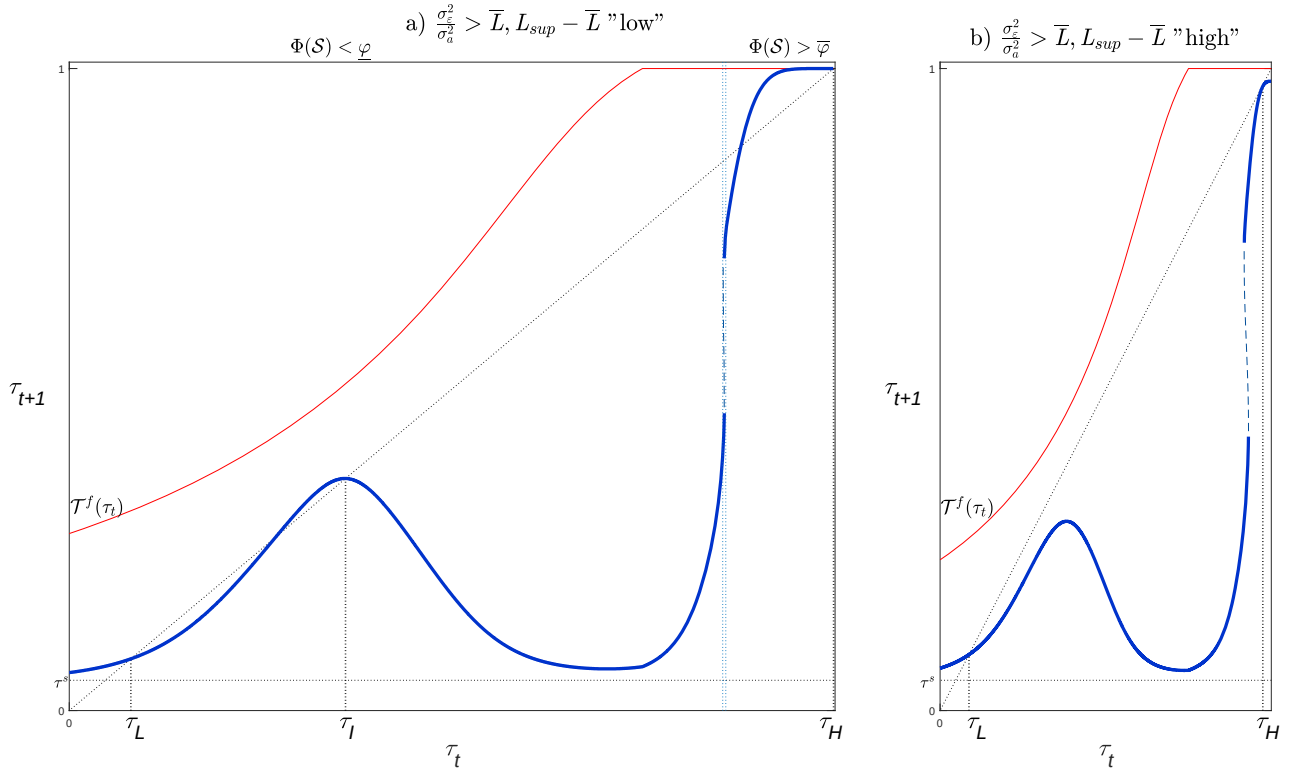

Figure 6: The long-term stability of three welfare-state regimes when $\frac{\sigma_\varepsilon^2}{\sigma_a^2} > \tilde{L}$ $\left(\frac{1}{2} < \gamma < 1\right.$; dashed line is for unstable beliefs$)$

4 Robustness with heterogenous distributive preferences

4.1 Distributive moral dilemma and preferences

If individual merit appears to be an important principle to characterize the level of redistribution that would be perceived as socially optimal, its content may give rise to interpretation and then be perceived differently from one individual to another. Certainly, effort and hard work are traditionally associated with individual merit. Certainly, luck is seen by most as an unfair component of income. However, as noted by Schokkaert and Truyts (2017), income differences caused by ability or talent may be seen more or less fair according to whether talent is perceived as reflecting former investments in human capital or as innate and then beyond an individual's control. This ambiguity is revealed by several studies whose findings are contradictory. For example, Fong (2001) and Corneo and Grüner (2002) show on American data that individuals who think that income is determined by luck rather than individual effort and ability are more favorable to redistribution. In the same line, Rustichini and Vostroknutov (2014) show in the lab that merit is attributed only if effort or ability affect the outcome. By contrast, Isaksson and Lindskog (2009) show that in Denmark individuals who believe that people get rewarded for their ability and talent are more favorable to redistribution. These contradictory findings may

suggest that the deservingness of income related to talent or ability can be perceived differently across individuals and across societies. In this line, comparing Norway to the United States, Almås et al. (2020) show that a larger share of the Americans adopt a *libertarian fairness* view (in which income inequality due to luck is considered fair) whereas a larger share of the Norvegians adopt an *egalitarian fairness* view (in which income inequality due to ability and talent is considered unfair), even if in both countries the standard *meritocratic fairness* view is most prevalent (37.5% in the United States, 42.5% in Norway).

In addition, individual merit cannot sum up all the principles of distributive justice (see Konow, 2003, for an overview). Forsé and Parodi (2006) show for example that European countries share an identical hierarchy of moral principles: first, the guarantee of basic needs; second, merit; and less important, equality of income. Besides, Durante et al. (2014) show that social concerns with respect to redistribution include both the concern for fairness and a dislike of inefficiency that can be associated with the "greatest aggregate happiness", i.e. the utilitarian concept of social justice most closely associated with Bentham (see Konow, 2003). These findings suggest that the principle characterized by "everyone should get what they deserve" can conflict with other moral concepts in defining a socially fair redistribution[20]. To the extent that the content of merit may be perceived differently from one individual to another and that different concepts of distributive justice may lead to the definition of different socially fair levels of redistribution, it is likely that individuals with similar information about the market outcomes will have different perceptions of the fair level of redistribution in the country.

In this optic, in addition to the meritocratic signal $\tau_t^{\hat{u}} = \arg\min_{\tau_t \in [0,1]} \left\{ \int_i \left(u_{it} - \hat{u}_{it} \right)^2 di \right\}$ which characterizes the principle *people should get what they deserve*, consider the utilitarian/efficiency signal $\tau_t^{\bar{u}} = \arg\max_{\tau_t \in [0,1]} \left\{ \int_i u_{it} di \right\}$ which defines the greater good for all whose relevance is supported in the literature to characterize social preferences (Charness and Rabin, 2002, Engelmann and Strobel, 2004, Fisman et al., 2007, 2017, Almås et al., 2020; see Fehr and Schmidt, 2006, for an overview). We then redefine the life-cycle utility function of an individual young at date $t-1$ as:

$$U_{it-1} = u_{it-1} - \frac{\varphi_{t-1}}{2} \left[\psi_i \left(\tau_t^{\hat{u}} - \tau_t \right)^2 + (1 - \psi_i) \left(\tau_t^{\bar{u}} - \tau_t \right)^2 \right] \tag{18}$$

where $\psi_i \in [0,1]$ is a parameter distributed across the population measuring the relative weight

[20] In recent years, a great deal of literature has showed experimentally that conflicts between deontological principles (such as the right to get what one deserves) and utilitarianism are a general feature of moral thinking (see Greene, 2008; Sinnott-Armstrong, 2008; Cushman and Young, 2009).

of the merit over efficiency in the distributive preferences[21]. Emphasizing the classical equity-efficiency trade-off, the social preferences as expressed in eq. (18) exhibit a standard form. We will assume that ψ_i of an individual of type i is independent of his intellectual efficiency a_i and of course of his luck or bad luck ε_i. For simplicity, we also assume that ψ_i is symetrically distributed such that the mean of the distribution equals its median: $\bar{\psi} = \psi_m$. As $\int_i u_{it} di = \frac{\bar{a}}{2}\left[1 - \gamma\left(\tau_t^e\right)^2 - (1-\gamma)\tau_t^2\right]$, it follows that (see Appendix E):

$$\tau_t^{\bar{u}} = \arg\max_{\tau\in[0,1]}\left\{\int_i u_{it} di\right\} = 0 \tag{19}$$

Obviously, as people are risk neutral in our setting, the optimal utilitarist taxation is nil. On the one hand, it may appear as an overly high simplification of the utilitarist view. On the other hand, results of Fehr et al. (2021) stress that risk aversion and the insurance motive play no role in the support for redistribution. Anyway, the important notion here is that people can have different views about the social optimal outcome, independently of their own social status.

4.2 Heterogeneity and multiplicity of the welfare-state regimes

As in the previous section, we can define the level of redistribution perceived as socially optimal by each individual as the optimal taxation level he would have chosen if the strength of the demand for social justice is infinitely high: $\tau_{it}^f = \lim_{\varphi_{t-1}\to\infty} \arg\max_{\tau\in[0,1]}\{U_{it}\}$. As each individual of type i has potentially a personal view τ_{it}^f of the social optimal tax rate, the social distance between the perceived social optimal tax rate and the chosen one at date t is now measured as:

$$\mathcal{S}_t = \int_i \left[\tau_{it}^f - \tau_t\right]^2 di = \sigma_\psi^2 \boldsymbol{\mathcal{T}^f}\left(\tau_t\right)^2 + \left[\bar{\psi}\boldsymbol{\mathcal{T}^f}\left(\tau_t\right) - \tau_t\right]^2 \tag{20}$$

where σ_ψ^2 denotes the variance of ψ. Let $\bar{\tau}^f = \bar{\psi}\boldsymbol{\mathcal{T}^f}(\tau)$ and $\sigma_{\tau^f}^2 = \sigma_\psi^2 \boldsymbol{\mathcal{T}^f}(\tau)^2$ denote respectively the mean and the variance of τ_i^f. $\sigma_{\tau^f}^2$ provides a measure of the degree of collective agreement to define $\bar{\tau}^f$ as the norm of fair redistribution. If $\sigma_{\tau^f}^2 = 0$, the perception of $\bar{\tau}^f$ as the fair redistribution is unanimously shared in the society. This is the case studied in section 3. By contrast, if $\sigma_{\tau^f}^2$ is high, $\bar{\tau}^f$ is of low significance in the population for defining a shared norm of fair level of redistribution. Social consensus declines. Therefore, everything else being equal, the social distance between the perceived social optimal tax rate grows with the variance of τ_i^f.

[21] In our setting all voters are equally concerned for others. Therefore, the heterogeneity we examine is different from the one in Dhami and al-Nowaihi (2010) in which a mixture of fair and selfish voters is considered.

How does it impact our results? If the distributions of a_i and τ_i^f are both symmetrical, the pivotal voter is the individual with the mean talent $\bar{a}$ and the mean perception $\bar{\tau}^f$ (see Appendix F). The dynamics of redistribution is then written exactly as in eq. (17), except that $\tau^s = 0$[22] and that the fair level of redistribution previously unanimously shared is replaced by its mean $\bar{\tau}^f$. Following the different Propositions highlighted in this article, the guarantee to obtain multiple stable stationary states with stable beliefs requires then that the concern for fairness be sufficiently high when the redistributive institution approaches the level reflecting the collective norm of fairness $\bar{\tau}^f$. Therefore, as the diversity of perceptions measured by $\sigma^2_{\tau^f}$ increases the social distance to distributive justice and then lowers the strength of the concern for fairness, a too high $\sigma^2_{\tau^f}$ may prevent the existence of multiple stable stationary states. The existence of multiple stationary states driven by our mechanism does not rely on the assumption that the perception of the fair redistributive tax is unanimously shared in the population, i.e. $\sigma^2_{\tau^f} = 0$, even if the social consensus to define $\bar{\tau}^f$ as the norm of fairness must be high enough. As $\sigma^2_{\tau^f}$ increases, the high-redistribution welfare-state regime first vanishes, then the intermediate, so that only the low-redistribution state remains.

5 Conclusion

After the Esping-Andersen' (1990) seminal study, welfare states are standardingly clustered in three identifiable regimes. In the liberal welfare regime, archetype of Anglo-Saxon countries, support is targeted at the poor and flat-rate benefits are low. In contrast, in the social-democratic welfare regime, archetype of Nordic countries, benefits are universal (available to all citizens) and high. Finally, in the corporatist welfare regime, archetype of Continental Europe, benefits are also high but linked to contributions, such that income redistribution is lower than in the social-democratic regime but greater than in the liberal one. By finding that most European continental countries are currently clustered in the high-taxation group along with Nordic countries, a recent study by Péligry and Ragot (2022) has suggested that the welfare states can evolve and change over time, casting doubt on the long-term stability of the canonical clustering. To study this issue, we have developped an overlapping generations model that merges the fairness approach of Alesina and Angeletos (2005) and the mechanism for the cultural transmission of the strength of moral norm proposed by Le Garrec (2018). Including these two components of the demand for redistribution, we identify a welfare-state clustering à la Esping-Andersen (1990) in which we assimilate a regime to a stable stationary state with perfect stable

[22] In that case, $\tau^s = 0$ and the model can no longer exhibit the Meltzer-Richard (1981) effect.

expectations. Three configurations with three regimes may arise depending on how unfair the income distribution is perceived (determined by luck rather than effort). In the one associated with the least unfairness, only the low-redistribution regime is truly stable. The two others, while responding to the definition of a stable stationary state with perfect stable expectations, can be destabilized by a self-fulfilling belief. They are only locally determinate. However, the clustering in three regimes is supposed to persist over time, even if the sets of countries composing the intermediate and the high-redistribution regimes can change. In the configuration associated with intermediate unfairness, only the intermediate-redistribution regime remains only locally determinate. This configuration predicts that countries with an intermediate welfare-state regime will end (in an indeterminate time) with a low-redistribution one, and the final clustering will encounter only two regimes, the low- and the high-redistribution. Finally, it is only in the configuration associated with the greatest unfairness that the clustering of the welfare states in three regimes is truly stable, with the sets of countries composing the different models unchanged over time. These results have been shown to be robust for extended specifications of the perception of the fair level of taxation incorporating heterogeneity across individuals.

As a first consequence of the specifications that we have used, through higher work effort market outcomes appear fairer in the US than in European countries. However, nothing seems to support the popular belief that American society is more socially mobile than European societies. Findings of Björklund and Jäntti (1997), Bratberg et al. (2017) and Helsø (2021) even support the view that intergenerational income mobility in Scandinavian countries is higher than in the United States. Second, we have considered socialization only through passive observation and imitation of the society at large (oblique socialization), and not through active efforts of the parents to transmit their values (vertical socialization). As mentioned, by assuming highly homogenous values at the regional level in most of the article, oblique socialization is highly efficient so that parents have few incentives to transmit their values. However, in the last section, we have stressed that individuals could use different concepts of distributive justice to define their fair level of redistribution. From these perspectives, incorporating both biased beliefs about social mobility and vertical socialization in the present analysis appears to be a promising avenue for further research.

References

[1] Ackert L., Martinez-Vazquez J. and Rider M. (2007), Social preferences and tax policy design: some experimental evidence, *Economic Inquiry*, 45(3), pp. 487-501.

[2] Akerlof G. (1997), Social distance and social decisions, *Econometrica*, 65(5), pp. 1005-1027.

[3] Alesina A., Glaeser E. and Sacerdote B. (2001), Why doesn't the US have a European-style welfare system?, *Brookings Papers on Economic Activity*, 2, pp. 187-277.

[4] Alesina A. and Angeletos G.-M. (2005), Fairness and redistribution: US versus Europe, *American Economic Review,* 95(4), pp. 960-980.

[5] Alesina A. and La Ferrara E. (2005), Preferences for redistribution in the land of opportunities, *Journal of Public Economics*, 89(5-6), pp. 897-931.

[6] Alesina A. and Fuchs-Schündeln N. (2007), Good bye Lenin (or not?): the effect of communism on people's preferences, *American Economic Review*, 97(4), pp. 1507-1528.

[7] Alesina A. and Giuliano P. (2011), Preferences for redistribution, in A. Bisin, J. Benhabib and M. Jackson eds. *Handbook of Social Economics*, North Holland Amsterdam, chap. 4, pp. 93-131.

[8] Alesina A., Algan Y., Cahuc P. and Giuliano P. (2015), Family values and the regulation of labor, *Journal of the European Economic Association*, 13(4), pp. 599-630.

[9] Algan Y. and Cahuc P. (2010), Inherited trust and growth, *American Economic Review*, 100(5), pp. 2060-2092.

[10] Almås I., Cappelen A. and Tungodden B. (2020), Cutthroat capitalism versus cuddly socialism: are Americans more meritocratic and efficiency-seeking than Scandinavians?, *Journal of Political Economy*, 128(5), pp. 1753-1788.

[11] Amable B. (2003), *The diversity of modern capitalism*, Oxford University Press.

[12] Bénabou R. and Tirole J. (2006), Belief in a just world and redistributive politics, *Quarterly Journal of Economics*, 121(2), pp. 699-746.

[13] Bisin A. and Verdier T. (2001), The economics of cultural transmission and the dynamics of preferences, *Journal of Economic Theory*, 97, pp. 298-319.

[14] Björklund A. and Jäntti M. (1997), Intergenerational income mobility in Sweden compared to the United States, *American Economic Review*, 87(5), pp. 1009-1018.

[15] Boldrin M. and Montes A. (2005), The intergenerational State education and pensions, *Review of Economic Studies*, 72(3), pp. 651-664.

[16] Bolton G. and Ockenfels A. (2000), ERC: A theory of equity, reciprocity, and competition, *American Economic Review*, 90(1), pp. 166-193.

[17] Boyd R. and Richerson P.J. (1985), *Culture and the evolutionary process*, London: University of Chicago Press.

[18] Boyd R., Richerson P.J. and Henrich J. (2011), The cultural niche: Why social learning is essential for human adaptation, *PNAS*, 108(suppl. 2), pp. 10918-10925.

[19] Bratberg E., Davis J., Mazumder B., Nybom M., Schnitzlein D. and Vaage K. (2017), A comparison of intergenerational mobility curves in Germany, Norway, Sweden, and the US, *Scandinavian Journal of Economics*, 119(1), pp. 72–101.

[20] Carroll C., Rhee B.-K. and Rhee C. (1994), Are there cultural effects on saving? Some cross-sectional evidence, *Quarterly Journal of Economics*, 109(3), pp. 685-699.

[21] Charness G. and Rabin M. (2002), Understanding social preferences with simple tests, *Quarterly Journal of Economics*, 117(3), pp. 817-869.

[22] Corneo G. (2001), Inequality and the State: Comparing US and German preferences, *Annals of Economics and Statistics*, 63/64, pp. 283-296.

[23] Corneo G. and Grüner H.-P.(2002), Individual preferences for political redistribution, *Journal of Public Economics*, 83, pp. 83-107.

[24] Corneo G. and Fong C. (2008), What's the monetary value of distributive justice?, *Journal of Public Economics*, 92(1), pp. 289-308.

[25] Corneo G. and Neher F. (2015), Democratic redistribution and rule of the majority, *European Journal of Political Economy*, 40, pp. 96-109.

[26] Cushman F. A. and Young L. (2009), The psychology of dilemmas and the philosophy of morality, *Ethical Theory and Moral Practice*, 12(1), pp. 9–24.

[27] Dhami S. and al-Nowaihi A. (2010), Redistributive policies with heterogeneous social preferences of voters, *European Economic Review*, 54, pp. 743-759.

[28] Docquier F., Paddison O. and Pestieau P. (2007), Optimal accumulation in an endogenous growth setting with human capital, *Journal of Economic Theory*, 134, pp. 361-378.

[29] Dohmen T., Falk A., Huffman D. and Sunde U. (2012), The intergenerational transmission of risk and trust attitudes, *Review of Economic Studies*, 79, pp. 645-677.

[30] Durante R., Putterman L. and van der Weele J. (2014), Preferences for redistribution and perception of fairness: an experimental study, *Journal of the European Economic Association*, 12(4), pp. 1059-1086.

[31] Engelmann D. and Strobel M. (2004), Inequality aversion, efficiency, and maximin preferences in simple distribution experiments, *American Economic Review*, 94(4), pp. 857-869.

[32] Enke B., Rodríguez-Padilla R. and Zimmermann F. (2022), Moral universalism and the structure of ideology, *Review of Economic Studies*, forthcoming.

[33] Esping-Andersen G. (1990), *The three worlds of welfare capitalism*, Polity press.

[34] Fehr E., Epper T. and Senn J. (2021), Other-regarding preferences and redistributive politics, Working Paper, University of Zurich, December.

[35] Fehr E. and Schmidt K. (2006), The economics of fairness, reciprocity and altruism: experimental evidence and New Theories, in S.-C. Kolm and J. Mercier Ythier (Eds), *Handbook of the economics of giving, altruism and reciprocity, vol. 1*, North Holland/Elsevier, chap. 8.

[36] Fernández R. and Fogli A. (2006), Fertility: the role of culture and family experience, *Journal of the European Economic Association*, 4(2–3) pp. 552–561.

[37] Ferrara M. (1996), The 'Southern model' of welfare in social Europe, *Journal of European Social Policy*, 6(1), pp. 17-37.

[38] Fisman R., Jakiela P. and Kariv S. (2017), Distributional preferences and political behavior, *Journal of Public Economics*, 155, pp. 1-10.

[39] Fisman R., Kariv S. and Markovits D. (2007), Individual preferences for giving, *American Economic Review*, 97(5), pp. 1858-1876.

[40] Fong C. (2001), Social preferences, self-interest, and the demand for redistribution, *Journal of Public Economics*, 82(2), pp. 225-246.

[41] Forsé M. and Parodi M. (2006), Justice distributive: la hiérachie des principes selon les européens, *Revue de l'OFCE*, 98, pp. 213-244.

[42] Funk P. (2005), Governemental action, social norms and criminal behavior, *Journal of Institutional and Theoretical Economics*, 161(3), pp. 522-535.

[43] Goodman R. and Peng I. (1996), The East Asian welfare states: Peripatetic learning, adaptive change, and nation-building, in Gøsta Esping-Andersen (ed.), *Welfare States in Transition: National Adaptations in Global Economies*, London: Sage, pp. 192–224.

[44] Greene J. (2008), The secret joke of Kant's soul, in W. Sinnott-Armstrong (ed.), *Moral psychology* (vol. 3), Cambridge, MA: MIT Press, pp. 35–80.

[45] Guiso L., Sapienza P. and Zingales L. (2006), Does culture affect economic outcomes?, *Journal of Economic Perspectives*, 20(2), pp. 23-48.

[46] Helsø A.-L. (2021), Intergenerational income mobility in Denmark and the United States, *Scandinavian Journal of Economics*, 123(2), pp. 508–531.

[47] Isaksson A.-S. and Lindskog A. (2009), Preferences for redistribution - A country comparison of fairness judments, *Journal of Economic Behavior & Organisation*, 72, pp. 884-902.

[48] Kerschbamer R. and Müller D. (2020), Social preferences and political attitudes: An online experiment on a large heterogeneous sample, *Journal of Public Economics*, 143.

[49] Konow J. (2003), Which is the fairest one of all? A positive analysis of justice theories, *Journal of Economic literature*, 41, pp. 1188-1239.

[50] Lefgren L., Sims D. and Stoddard O. (2016), Effort, luck and voting for redistribution, *Journal of Public Economics*, 182, pp. 89-97.

[51] Le Garrec G. (2018), Fairness, social norms and the cultural demand for redistribution, *Social Choice and Welfare*, 50(2), pp. 191-212.

[52] Le Grand F., Ragot X. and Rodrigues D. (2022), The welfare of nations: do social preferences matter for the macroeconomy?, mimeo, April.

[53] Lindbeck A., Nyberg S. and Weibull J. (1999), Social norms and economic incentives in the welfare state, *Quarterly Journal of Economics*, 114(1), pp. 1-35.

[54] Luttmer E. and Singhal M. (2011), Culture, context, and the taste for redistribution, *American Economic Journal: Economic Policy*, 3(1), pp. 157-179.

[55] Malmendier U. and Nagel S. (2011), Depression babies: do macroeconomic experience affect risk taking?, *Quarterly Journal of Economics*, 126, pp. 373-416.

[56] McCrae R. and Costa P. (1994), The stability of personality: observation and evaluations, *Current Directions in Psychological Science*, 3(6), pp. 173-175.

[57] Meltzer A. and Richard S. (1981), A rational theory of the size of government, *Journal of Political Economy*, 89(5), pp 914-927.

[58] Neundorf A. and Smets K. (2017), Political socialization and the making of citizens, in *Oxford Handbooks Online in Political Science*. Oxford University Press.

[59] Péligry P. and Ragot X. (2022), Evolution of fiscal systems: convergence or divergence?, Sciences Po OFCE Working Paper n°3/2022.

[60] Persson T. and Tabellini G. (2021), Culture, institutions, and policy, in A. Bisin and G. Federico (Eds), *Hanbook of Historical Economics*, Elsevier Science, chap. 16, pp. 463-489.

[61] Piketty T. (1995), Social mobility and redistributive politics, *Quarterly Journal of Economics*, 110(3), pp. 551-584.

[62] Postlewaite A. (2011), Social norms and preferences, in A. Bisin, J. Benhabib and M. Jackson eds. *Handbook of Social Economics*, North Holland Amsterdam, chap. 2, pp. 31-67.

[63] Roberts B. and DelVecchio W. (2000), The rank-order consistency of personality traits from childhood to old age: a quantitative review of longitudinal studies, *Psychological Bulletin*, 126(1), pp. 3-25.

[64] Roland G. (2020), The deep historical roots of modern culture: A comparative perspective, *Journal of Comparative Economics*, 48, pp. 483-508.

[65] Roth C. and Wohlfart J. (2018), Experienced inequality and preferences for redistribution, *Journal of Public Economics*, 167, pp. 251-262.

[66] Rustichini A. and Vostroknutov A. (2014), Merit and Justice: An expirimental analysis of attitude to inequality, *PLoS ONE* 9(12):e114512, doi:10.1371/journal.pone.0114512.

[67] Sands M. (2017), Exposure to inequality affects support for redistribution, *Proceedings of the National Academy of Sciences*, 114(4), pp. 663-668.

[68] Schildberg-Hörisch H. (2010), Is the veil of ignorance only a concept about risk? An experiment, *Journal of Public Economics*, 94, pp. 1062-1066.

[69] Schokkaert E. and Truyts T. (2017), Preferences for redistribution and social structure, *Social Choice and Welfare*, 49, pp. 545-576.

[70] Sinnott-Armstrong W. (2008), Abstract+concrete=paradox, in S. Nichols & J. Knobe (Eds.), *Experimental philosophy*, NY: Oxford University Press, pp. 209–230.

[71] Tabellini G. (2008), The scope of cooperation: values and incentives, *Quarterly Journal of Economics*, 123, pp. 905-950.

[72] Tyran J.-R. and Sausgruber (2006), A little fairness may induce a lot of redistribution in democracy, *European Economic Review*, 50, pp. 469-485.

[73] Weber, M. (1905), *The Protestant ethic and the spirit of capitalism* (German: Die protestantische ethik und der geist des kapitalismus).

Appendix A. Proof of Lemma 1

By definition, $\tau_t^f = \arg\min_{\tau} \int_i (u_{it} - \hat{u}_{it})^2 di$. When considering eqs. (2), (5), (7), (8) and (9) it yields:

$$\begin{aligned} u_{it} &= (a_i [1 - \gamma\tau_t^e - (1-\gamma)\tau_t] + \varepsilon_{it})(1-\tau_t) + \tau_t \bar{a}[1 - \gamma\tau_t^e - (1-\gamma)\tau_t] \\ &\quad -\frac{a_i}{2}\left[\gamma(1-\tau_t^e)^2 + (1-\gamma)(1-\tau_t)^2\right] \end{aligned} \tag{21}$$

Knowing that $\hat{u}_{it} = \hat{y}_{it} - \frac{1}{2\beta_i}\left[\gamma h_{it-1}^2 + (1-\gamma)e_{it}^2\right]$ where $\hat{y}_{it} = A_i[\gamma h_{it-1} + (1-\gamma)e_{it}]$, it follows from eq. (7) and (8) that:

$$\hat{u}_{it} = a_i[1 - \gamma\tau_t^e - (1-\gamma)\tau_t] - \frac{a_i}{2}\left[\gamma(1-\tau_t^e)^2 + (1-\gamma)(1-\tau_t)^2\right] \tag{22}$$

Therefore, from eqs (21) and (22) it yields $u_{it} - \hat{u}_{it} = (1-\tau_t)\varepsilon_{it} - \tau_t[1 - \gamma\tau_t^e - (1-\gamma)\tau_t](a_i - \bar{a})$, and as luck ε and personal talent a are independently distributed, it follows:

$$\int_i (u_{it} - \hat{u}_{it})^2 di = (1-\tau_t)^2\sigma_\varepsilon^2 + \tau_t^2[1 - \gamma\tau_t^e - (1-\gamma)\tau_t]^2\sigma_a^2$$

It follows that $\tau^f = \mathcal{T}^f(\tau^e) \equiv \arg\min_{\tau\in[0,1]}\left\{(1-\tau)^2\sigma_\varepsilon^2 + \tau^2[1 - \gamma\tau^e - (1-\gamma)\tau]^2\sigma_a^2\right\}$.

As is obvious, $\lim_{\frac{\sigma_\varepsilon^2}{\sigma_a^2}\to 0^+} \mathcal{T}^f(\tau^e) = 0$ (if $\tau^e < 1$) and $\lim_{\frac{\sigma_\varepsilon^2}{\sigma_a^2}\to +\infty} \mathcal{T}^f(\tau^e) = 1$. The first order condition $\frac{\partial\left\{(1-\tau)^2\sigma_\varepsilon^2 + \tau^2[1-\gamma\tau^e-(1-\gamma)\tau]^2\sigma_a^2\right\}}{\partial\tau} = 0$ rewrites as:

$$-\left(1-\tau^f\right)\frac{\sigma_\varepsilon^2}{\sigma_a^2} + \tau^f\left[1 - \gamma\tau^e - (1-\gamma)\tau^f\right]^2 - (1-\gamma)\tau^{f2}\left[1 - \gamma\tau^e - (1-\gamma)\tau^f\right] = 0 \tag{23}$$

The second order condition is then $\frac{\partial^2\left\{(1-\tau)^2\sigma_\varepsilon^2 + \tau^2[1-\gamma\tau^e-(1-\gamma)\tau]^2\sigma_a^2\right\}}{\partial\tau^2} \geq 0$, or equivalently:

$$\frac{\sigma_\varepsilon^2}{\sigma_a^2} + 6(1-\gamma)^2\tau^{f2} - 6(1-\gamma)(1-\gamma\tau^e)\tau^f + (1-\gamma\tau^e)^2 \geq 0 \tag{24}$$

Differentiating eq. (23) leads to
$\frac{\partial^2\left\{(1-\tau)^2\sigma_\varepsilon^2 + \tau^2[1-\gamma\tau^e-(1-\gamma)\tau]^2\sigma_a^2\right\}}{\partial\tau^2}d\tau^f + \frac{\partial^2\left\{(1-\tau)^2\sigma_\varepsilon^2 + \tau^2[1-\gamma\tau^e-(1-\gamma)\tau]^2\sigma_a^2\right\}}{\partial\tau\partial\tau^e}d\tau^e = 0$
where $\frac{\partial^2\left\{(1-\tau)^2\sigma_\varepsilon^2 + \tau^2[1-\gamma\tau^e-(1-\gamma)\tau]^2\sigma_a^2\right\}}{\partial\tau\partial\tau^e} = \gamma\tau[3(1-\gamma)\tau - 2(1-\gamma\tau^e)]$. With condition (24), it follows that:

$$\frac{d\tau^f}{d\tau^e} \geq 0 \Leftrightarrow \tau^f \leq \frac{2(1-\gamma\tau^e)}{3(1-\gamma)}$$

Considering the first order condition (23), if $\tau^e = 2 - \frac{1}{\gamma}$ or $\tau^e = 1$ it yields $\mathcal{T}^f(\tau^e) = 1$, where $\left.\frac{d\tau^f}{d\tau^e}\right|_{\tau^f=\tau^e=1} = -\frac{\gamma(1-\gamma)}{\frac{\sigma_\varepsilon^2}{\sigma_a^2}+(1-\gamma)^2} \leq 0$ and $\left.\frac{d\tau^f}{d\tau^e}\right|_{\tau^f=1,\tau^e=2-\frac{1}{\gamma}} = \frac{\gamma(1-\gamma)}{\frac{\sigma_\varepsilon^2}{\sigma_a^2}-2(1-\gamma)^2} \geq 0$.

Considering then the second order condition (24), it yields that $\left.\frac{\partial^2\left\{(1-\tau)^2\sigma_\varepsilon^2+\tau^2[1-\gamma\tau^e-(1-\gamma)\tau]^2\sigma_a^2\right\}}{\partial\tau^2}\right|_{\tau^f=1,\tau^e=2-\frac{1}{\gamma}} \geq 0$ is equivalent to

$$\frac{\sigma_\varepsilon^2}{\sigma_a^2} \geq 2\left(1-\gamma\right)^2$$

where $\frac{\partial\left.\frac{\partial^2\left\{(1-\tau)^2\sigma_\varepsilon^2+\tau^2[1-\gamma\tau^e-(1-\gamma)\tau]^2\sigma_a^2\right\}}{\partial\tau^2}\right|_{\tau=1}}{\partial\tau^e} = 6\gamma\left(1-\gamma\right)\left[1-\frac{1}{3}\frac{1-\gamma\tau^e}{1-\gamma}\right] \geq 0\ \forall\tau^e \geq 2-\frac{1}{\gamma}$.

As $\left(\tau^f,\tau^e\right)\in[0,1]^2$, it yields:

1. if $\gamma>\frac{1}{2}$ and $\frac{\sigma_\varepsilon^2}{\sigma_a^2}\geq 2\left(1-\gamma\right)^2$, $\tau^e\geq 2-\frac{1}{\gamma}$ yields $\mathcal{T}^f\left(\tau^e\right)=1$,
2. if $\gamma\leq\frac{1}{2}$ and $\frac{\sigma_\varepsilon^2}{\sigma_a^2}\geq\frac{1}{2}$, $\mathcal{T}^f\left(\tau^e\right)=1\ \forall\tau^e$.

Summerizing the above points (1) and (2), one gets: if $\frac{\sigma_\varepsilon^2}{\sigma_a^2}\geq\min\left\{2\left(1-\gamma\right)^2,\frac{1}{2}\right\}$ there exists $\tilde{\tau}=\max\left\{2-\frac{1}{\gamma},0\right\}$ so that $\tau^e\geq\tilde{\tau}$ yields $\mathcal{T}^f\left(\tau^e\right)=1$. Thereafter, when considering $\gamma>\frac{1}{2}$ and $\tau^e<\tilde{\tau}$, differentiating eq. (23) yields:

- $\frac{\partial\mathcal{T}^f}{\partial\tau^e}\left(\tau^e\right)\geq 0$ as $\tau^f\leq\frac{2(1-\gamma\tau^e)}{3(1-\gamma)}\left(>1\ \forall\tau^e<\tilde{\tau}\right)$, where $\lim_{\gamma\to 1}\frac{\partial\mathcal{T}^f}{\partial\tau^e}\left(0\right)=\frac{2\frac{\sigma_\varepsilon^2}{\sigma_a^2}}{\left(1+\frac{\sigma_\varepsilon^2}{\sigma_a^2}\right)^2}\leq\frac{1}{2}$,
- $\frac{\partial\mathcal{T}^f}{\partial\frac{\sigma_\varepsilon^2}{\sigma_a^2}}\left(\tau^e\right)\geq 0$ with $\lim_{\frac{\sigma_\varepsilon^2}{\sigma_a^2}\to+\infty}\mathcal{T}^f\left(\tau^e\right)=1$,
- $\frac{\partial\mathcal{T}^f}{\partial\gamma}\left(0\right)=\frac{\mathcal{T}^f(0)^2\left[4(1-\gamma)\mathcal{T}^f(0)-3\right]}{\frac{\partial^2\left\{(1-\tau)^2\sigma_\varepsilon^2+\tau^2[1-\gamma\tau^e-(1-\gamma)\tau]^2\sigma_a^2\right\}}{\partial\tau^2}}<0$,

According to eq. (23), resolving equation $\mathcal{T}^f\left(\tau\right)=\tau$ leads to $-\left(1-\tau\right)\frac{\sigma_\varepsilon^2}{\sigma_a^2}+\tau\left[1-\tau\right]^2-\left(1-\gamma\right)\tau^2\left[1-\tau\right]=0$. $\tau=1$ is one obvious root. The others are associated with the following equation $-\left(2-\gamma\right)\tau^2+\tau-\frac{\sigma_\varepsilon^2}{\sigma_a^2}=0$ that has no real root unless $\frac{\sigma_\varepsilon^2}{\sigma_a^2}\leq\frac{1}{4(2-\gamma)}$.

Appendix B. Proof of Proposition 2

When considering equilibrium, dynamics (12) rewrites as $\frac{\mathcal{T}^{\boldsymbol{f}}(\tau)-\tau}{\mathcal{T}^{\boldsymbol{f}}(\tau)-\tau^s} = \frac{\gamma\Delta+2(1-\gamma)(1+\Delta)}{\gamma\Delta+2(1-\gamma)(1+\Delta)+\varphi}$, where $\tau^s = \frac{\Delta}{\gamma\Delta+2(1-\gamma)(1+\Delta)}$ (eq. 13).

- $\underline{\varphi}$ is the lower value of φ such that the basic model exhibits two stable equilibria. For low values of $\frac{\sigma_\varepsilon^2}{\sigma_a^2}$, $\underline{\varphi}$ is associated with the threshold point $\tilde{\tau} = 2 - \frac{1}{\gamma}$, $\frac{1}{2} < \gamma < 1$, beyond which the perceived fair tax rate τ^f is equal to 1 (Lemma 1). Accordingly, $\underline{\varphi}$ must solve eq. $\frac{1-\tilde{\tau}}{1-\tau^s} = \frac{\gamma\Delta+2(1-\gamma)(1+\Delta)}{\gamma\Delta+2(1-\gamma)(1+\Delta)+\varphi}$, such that $\underline{\varphi} = \underline{\varphi}^{cst} = \frac{\tilde{\tau}-\tau^s}{1-\tilde{\tau}}\left[\gamma\Delta + 2\left(1-\gamma\right)\left(1+\Delta\right)\right]$ is constant. For higher values of $\frac{\sigma_\varepsilon^2}{\sigma_a^2}$, in $\varphi = \underline{\varphi}^{cst}$, eq. $\frac{\mathcal{T}^{\boldsymbol{f}}(\tau)-\tau}{\mathcal{T}^{\boldsymbol{f}}(\tau)-\tau^s} = \frac{\gamma\Delta+2(1-\gamma)(1+\Delta)}{\gamma\Delta+2(1-\gamma)(1+\Delta)+\varphi}$ still exhibits three equilibria so that, by definition, $\underline{\varphi} < \underline{\varphi}^{cst}$ becomes slightly decreasing with $\frac{\sigma_\varepsilon^2}{\sigma_a^2}$ (see Figure 6).

- $\overline{\varphi}$ is the higher value of φ such that the basic model exhibits two stable equilibria when $\frac{\sigma_\varepsilon^2}{\sigma_a^2} > \tilde{L}$, i.e. when the graph of $\mathcal{T}^{\boldsymbol{f}}(\tau)$ cuts the main diagonal only once in $\tau = 1$, i.e. $\mathcal{T}^{\boldsymbol{f}}(\tau) > \tau \ \forall \tau < 1$ (Lemma 1). Indeed, as a large φ tends to bring the graph of τ as defined by eq. $\frac{\mathcal{T}^{\boldsymbol{f}}(\tau)-\tau}{\mathcal{T}^{\boldsymbol{f}}(\tau)-\tau^s} = \frac{\gamma\Delta+2(1-\gamma)(1+\Delta)}{\gamma\Delta+2(1-\gamma)(1+\Delta)+\varphi}$ closer to the graph of $\mathcal{T}^{\boldsymbol{f}}(\tau)$, a too large φ implies there is only one equilibrium if $\frac{\sigma_\varepsilon^2}{\sigma_a^2} > \tilde{L}$. The tangency between the graph of τ and the main diagonal, if it is possible, is obtained in $\hat{\tau}$ where the distance between the perceived fair tax rate $\mathcal{T}^{f}(\tau)$ and the effective tax rate τ is minimal on the interval $[0, \tilde{\tau}]$. Define then $\hat{\tau} = \underset{\tau\in[0,\tilde{\tau}]}{\arg\min}\left[\mathcal{T}^{\boldsymbol{f}}(\tau) - \tau\right]^2$ for any $\frac{\sigma_\varepsilon^2}{\sigma_a^2} > \tilde{L}$, it follows that $\overline{\varphi} = \frac{\hat{\tau}-\tau^s}{\mathcal{T}^{\boldsymbol{f}}(\hat{\tau})-\hat{\tau}}\left[\gamma\Delta + 2\left(1-\gamma\right)\left(1+\Delta\right)\right]$. As the distance between $\mathcal{T}^{\boldsymbol{f}}(\hat{\tau})$ and $\hat{\tau}$ tends to 0^+ when $\frac{\sigma_\varepsilon^2}{\sigma_a^2} \to \tilde{L}^+$, $\overline{\varphi}$ becomes infinitly high in that case, $\lim_{\frac{\sigma_\varepsilon^2}{\sigma_a^2}\to\tilde{L}^+} \overline{\varphi} = +\infty$, and decreases with $\frac{\sigma_\varepsilon^2}{\sigma_a^2}$ (see Figure 7).

- Afterwards, L_{sup} is defined such that $\underline{\varphi}(L_{\text{sup}}) = \overline{\varphi}(L_{\text{sup}})$.

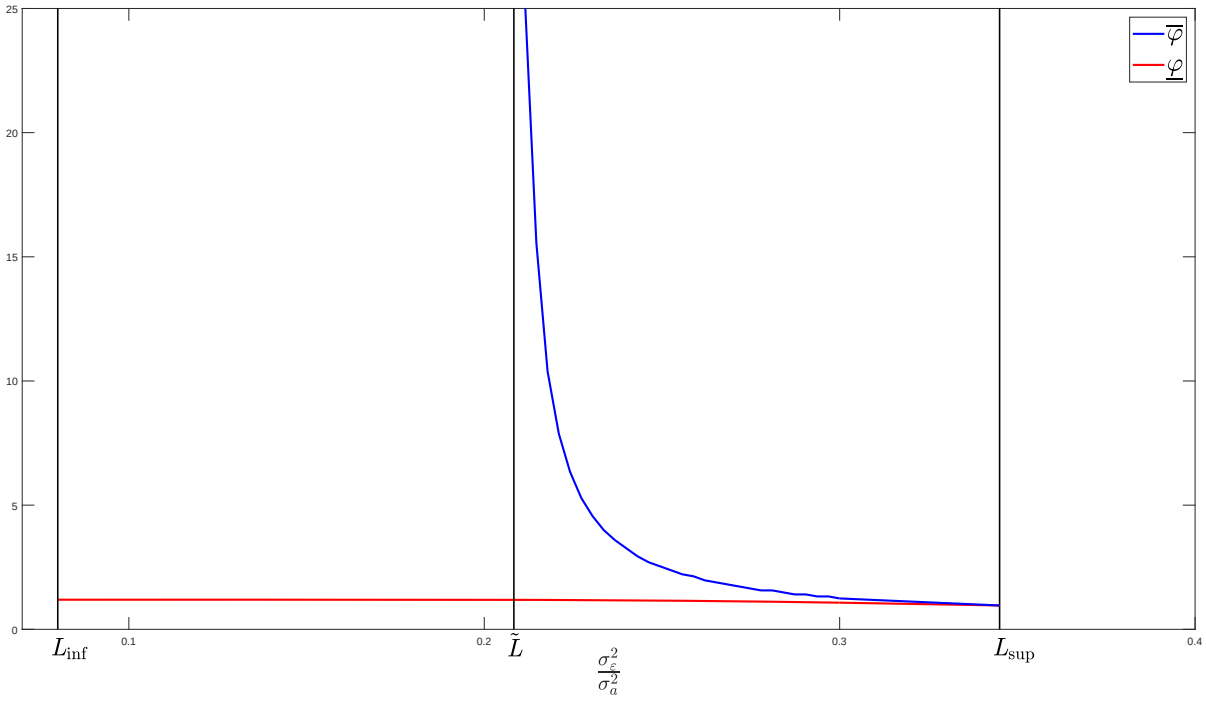

Figure 7: $\overline{\varphi}\left(\frac{\sigma_\varepsilon^2}{\sigma_a^2}\right)$ and $\underline{\varphi}\left(\frac{\sigma_\varepsilon^2}{\sigma_a^2}\right)$

Appendix C. Existence of τ_I and τ_{SD} conditional on $\Phi(0)$

At equilibrium, the forward dynamics (12) is rewritten as $\tau = \tilde{\xi}\tau^s + \left(1-\tilde{\xi}\right)\mathcal{T}^{\boldsymbol{f}}(\tau)$, where $\tilde{\xi} = \frac{\gamma\Delta+2(1-\gamma)(1+\Delta)}{\gamma\Delta+2(1-\gamma)(1+\Delta)+\varphi} \in (0,1]$. It follows that $\mathcal{T}^{\boldsymbol{f}}(\tau)-\tau = \tilde{\xi}\left(\mathcal{T}^{\boldsymbol{f}}(\tau)-\tau^s\right)$, i.e. the gap between the fair and the effective level of taxation is growing with the fair level of taxation.

Take now the complete dynamics (17). We know that on intervals $\left[0,\tau_1^f\right]$ and $\left[\tilde{\tau},\tau_3^f\right]$, it describes an increasing relationship between τ_{t+1} and τ_t: $\frac{d\tau_{t+1}}{d\tau_t} \geq 0$. In addition, if we set initially at period $t=0$ $\tau_0 = \tau_1^f$ or $\tau_0 = \tau_3^f$, τ_1^f and $\tau_3^f\,(=1)$ being two roots of eq. $\tau = \mathcal{T}^{\boldsymbol{f}}(\tau)$, $\Phi\left(\left[\mathcal{T}^{\boldsymbol{f}}(\tau_0)-\tau_0\right]^2\right) = \Phi\left(\left[\mathcal{T}^{\boldsymbol{f}}\left(\tau_1^f\right)-\tau_1^f\right]^2\right) = \Phi\left(\left[\mathcal{T}^{\boldsymbol{f}}\left(\tau_3^f\right)-\tau_3^f\right]^2\right) = \Phi(0) < +\infty$ and $\lim\limits_{\tau_t\to\tau_1^f}\frac{d\tau_{t+1}}{d\tau_t} = \lim\limits_{\tau_t\to\tau_3^f}\frac{d\tau_{t+1}}{d\tau_t} = 0$. It follows that the gap between the fair and the effective level of taxation conditionnally to $\tau_0 = \tau_3^f$ is higher than the gap conditionnally to $\tau_0 = \tau_1^f$ at period $t=1$ because $\tau_3^f > \tau_1^f$ (see above). It implies that $\Phi(0) > \Phi\left(\left[\mathcal{T}^{\boldsymbol{f}}(\tau_1)-\tau_1\right]^2\middle|\,\tau_0=\tau_1^f\right) > \Phi\left(\left[\mathcal{T}^{\boldsymbol{f}}(\tau_1)-\tau_1\right]^2\middle|\,\tau_0=\tau_3^f\right)$ and equivalently that $\tilde{\xi}_0 < \tilde{\xi}_1\left(\tau_0=\tau_1^f\right) < \tilde{\xi}_1\left(\tau_0=\tau_3^f\right)$. Starting from τ_0 satisfying $\tau_0 = \mathcal{T}^{\boldsymbol{f}}(\tau_0)$, the decrease in τ_t is steeper if $\tau_0=\tau_3^f$ than if $\tau_0=\tau_1^f$. In the (τ_t,τ_{t+1}) plane, superimposing τ_1^f and τ_3^f allows us to highlight three configurations:

- if $\Phi(0)$ is sufficiently high, two stationary states with tax rates denoted τ_I and τ_H in the text, close to τ_1^f and τ_3^f, respectively, exist such that $\Phi\left(\left[\mathcal{T}^{\boldsymbol{f}}(\tau_I)-\tau_I\right]^2\right) > \Phi\left(\left[\mathcal{T}^{\boldsymbol{f}}(\tau_H)-\tau_H\right]^2\right)$ (Fig. 8a),

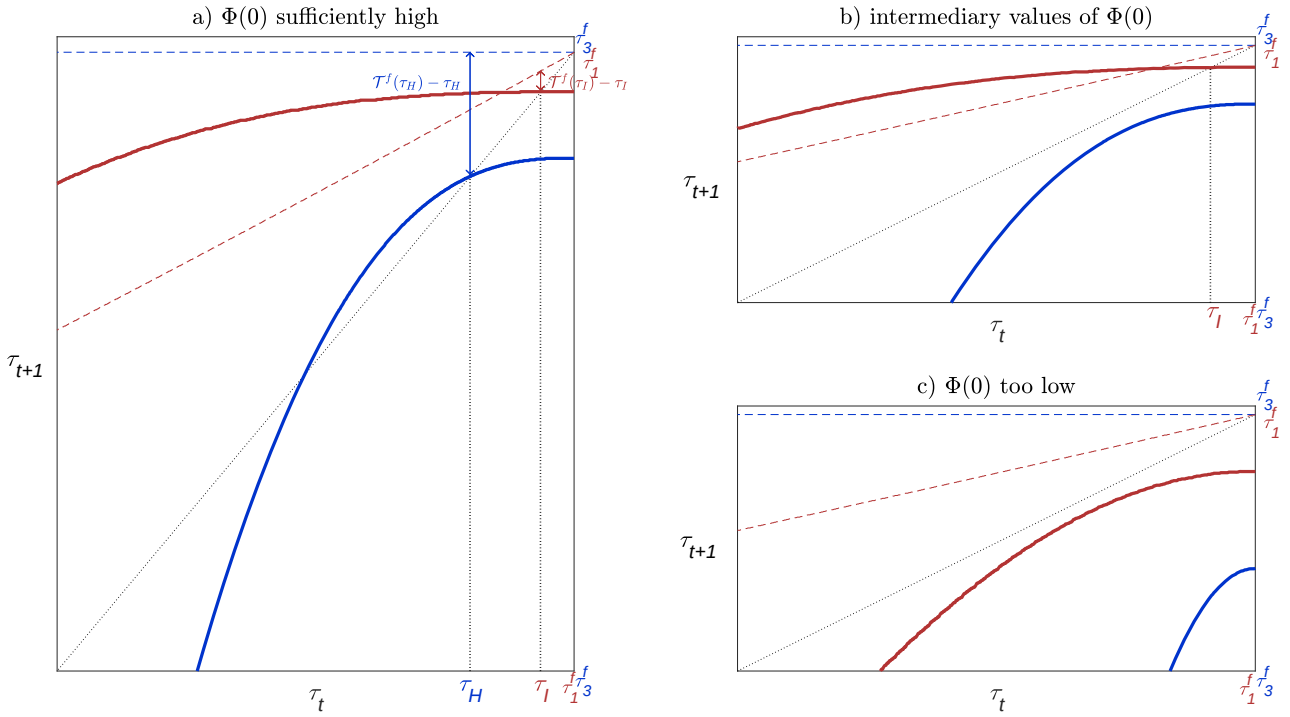

Figure 8: Existence of τ_I and τ_{SD} conditional on $\Phi(0)$ (τ_1^f and τ_3^f superimposed)

- if $\Phi(0)$ is too low both stationary states τ_I and τ_H vanish (Fig. 8c),
- for intermediate values of $\Phi(0)$ the stationary state τ_I exists while the stationary state τ_H vanishes (Fig. 8b).

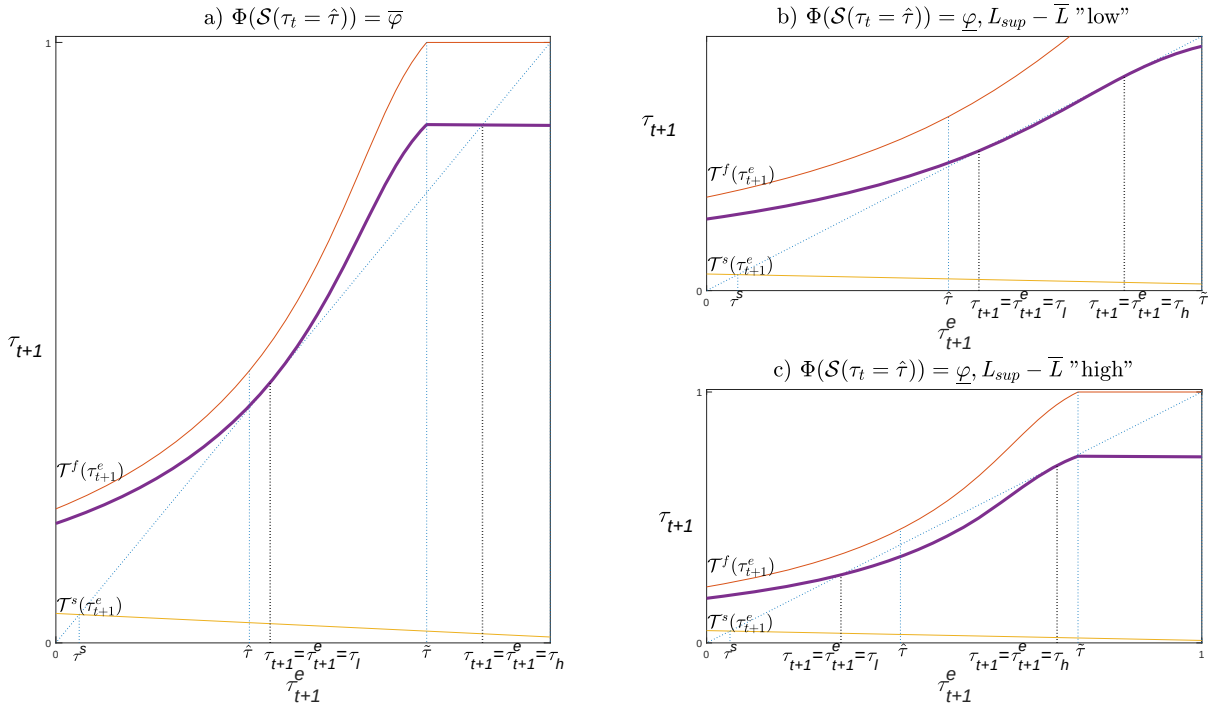

Figure 9: Comparing τ_l and $\hat{\tau}$ in $\varphi = \overline{\varphi}$ and $\varphi = \underline{\varphi}$

Appendix D. Comparing τ_l and $\hat{\tau}$ when $\frac{\sigma_\varepsilon^2}{\sigma_a^2} = \underline{L}$ and $\frac{\sigma_\varepsilon^2}{\sigma_a^2} = \overline{L}$

1. $\frac{\sigma_\varepsilon^2}{\sigma_a^2} = \underline{L}$ $(\Rightarrow \Phi(\mathcal{S}(\hat{\tau})) = \overline{\varphi}(\underline{L}))$.

   - In the $(\tau^e_{t+1}, \tau_{t+1})$ plane, assume that $\varphi = \varphi^*$ is such that the graph of $\tau_{t+1} = \xi(\varphi^*)\mathcal{T}^s(\tau^e_{t+1}) + (1 - \xi(\varphi^*))\mathcal{T}^f(\tau^e_{t+1})$ (eq. 12) crosses the main diagonal in $\tau^e_{t+1} = \hat{\tau}$, i.e. also $\tau_{t+1} = \hat{\tau}$. By definition of $\hat{\tau}$ $\frac{\partial \mathcal{T}^f}{\partial \tau}(\tau^e_{t+1} = \hat{\tau}) = 1$, and it follows that $\frac{d\tau_{t+1}}{d\tau^e_{t+1}}\Big|_{\tau_{t+1}=\tau^e_{t+1}=\hat{\tau}} = \xi(\varphi^*) + (1 - \xi(\varphi^*))\frac{\partial \mathcal{T}^s}{\partial \tau}(\tau^e_{t+1} = \hat{\tau}) < 1$. As $\mathcal{T}^s(\tau^e_{t+1}) < \mathcal{T}^f(\tau^e_{t+1})$ by assumption, $\frac{\partial \tau_{t+1}}{\partial \varphi} > 0$ and we derive that $\overline{\varphi} > \varphi^*$ exists such that $\tau_{t+1} = \tau^e_{t+1} = \{\tau_l, \tau_h\}$, where $\tau_h > \tau_l > \hat{\tau}$ and $\frac{d\tau_{t+1}}{d\tau^e_{t+1}}\Big|_{\tau_{t+1}=\tau^e_{t+1}=\tau_l} = 1$.
   - Thereafter, studying the case $\frac{\sigma_\varepsilon^2}{\sigma_a^2} = \underline{L}$ in which $\Phi(\mathcal{S}(\tau_t = \hat{\tau})) = \overline{\varphi}(\underline{L})$, it follows that $\tau_t = \hat{\tau} \Rightarrow \tau_{t+1} = \tau^e_{t+1} = \{\tau_l, \tau_h\}$, where $\tau_h > \tau_l > \hat{\tau}$ (Fig 9a).

2. $\frac{\sigma_\varepsilon^2}{\sigma_a^2} = \overline{L}$ $\left(\Rightarrow \Phi(\mathcal{S}(\hat{\tau})) = \underline{\varphi}(\overline{L})\right)$.

   - As $\underline{\varphi}\left(\frac{\sigma_\varepsilon^2}{\sigma_a^2}\right) \to \overline{\varphi}\left(\frac{\sigma_\varepsilon^2}{\sigma_a^2}\right)$ when $\frac{\sigma_\varepsilon^2}{\sigma_a^2} \to L_{\text{sup}}$ and $\tau_l > \hat{\tau}$ when $\varphi = \overline{\varphi}$ (see above), it follows that $\tau_l > \hat{\tau}$ in $\Phi(\mathcal{S}(\hat{\tau})) = \underline{\varphi}(\overline{L})$ when $L_{\text{sup}} - \overline{L}$ is sufficiently low (Fig. 9b).
   - By contrast, if $L_{\text{sup}} - \overline{L}$ is sufficiently high, $\underline{\varphi}(\overline{L})$ becomes low enough compared to $\overline{\varphi}(\overline{L})$ so that $\tau_l < \hat{\tau}$ in $\Phi(\mathcal{S}(\hat{\tau})) = \underline{\varphi}(\overline{L})$ (Fig. 9c).

Appendix E. The optimal utilitarist taxation

According to eqs. (2), (7) and (8), the private utility (5) can be rewritten as:

$$u_{it} = y_{it}\left(1-\tau_t\right) + \tau_t \bar{y}_t - \frac{a_i}{2}\left[\gamma\left(1-\tau_t^e\right)^2 + \left(1-\gamma\right)\left(1-\tau_t\right)^2\right]$$

Accordingly, the average private utility becomes:

$$\bar{u}_t = \int_i u_{it} di = \bar{y}_t - \frac{\bar{a}}{2}\left[\gamma\left(1-\tau_t^e\right)^2 + \left(1-\gamma\right)\left(1-\tau_t\right)^2\right]$$

Using eq. (9), it then yields:

$$\bar{u}_t = \frac{\bar{a}}{2}\left[1-\gamma\left(\tau_t^e\right)^2 - \left(1-\gamma\right)\tau_t^2\right]$$

It is then straightforward that $\tau_t^{\bar{u}} = \arg\max_{\tau_t \in [0,1]} \left\{\int_i u_{it} di\right\} = 0$.

Appendix F. Heterogenous views and the condorcet winner

If assuming that the level of taxation perceived as fair may differ across the population, $\tau_i^f \neq \tau_j^f \ \forall i \neq j$, the median voter is no more necessarily the condorcet winner. To see this, express eq. (11) as $\tau_i = \frac{(1-\gamma\tau^e)(\bar{a}-a_i)+\varphi\tau_i^f}{(1-\gamma)(2\bar{a}-a_i)+\varphi}$. For the median voter to be the condorcet winner at equilibrium, then the tax rate $\tau = \tau^e = \tau_m = \frac{(1-\gamma\tau_m)(\bar{a}-a_m)+\varphi\tau_m^f}{(1-\gamma)(2\bar{a}-a_m)+\varphi}$ must verify $\Pr[\tau_i \geq \tau] = \frac{1}{2}$. The condition $\tau_i \geq \tau_m$ corresponds to the set $\mathcal{C} = \left\{ \left(a_i, \tau_i^f\right) \ / \ \tau_i^f \geq \frac{\frac{(1-\gamma)(2\bar{a}-a_i)+\varphi}{(1-\gamma)(2\bar{a}-a_m)+\varphi}\left[(1-\gamma\tau^e)(\bar{a}-a_m)+\varphi\tau_m^f\right]-(1-\gamma\tau^e)(\bar{a}-a_i)}{\varphi} \right\}$. It follows that $\Pr[\tau_i \geq \tau] = \frac{1}{2} \Leftrightarrow \Pr\{\mathcal{C}\} = \frac{1}{2}$.

$\frac{(1-\gamma\tau^e)(\bar{a}-a_i)+\varphi\tau_i^f}{(1-\gamma)(2\bar{a}-a_i)+\varphi} \geq \frac{(1-\gamma\tau^e)(\bar{a}-a_m)+\varphi\tau_m^f}{(1-\gamma)(2\bar{a}-a_m)+\varphi}$

$\Leftrightarrow (1-\gamma\tau^e)(\bar{a}-a_i) + \varphi\tau_i^f \geq \frac{(1-\gamma)(2\bar{a}-a_i)+\varphi}{(1-\gamma)(2\bar{a}-a_m)+\varphi}\left[(1-\gamma\tau^e)(\bar{a}-a_m) + \varphi\tau_m^f\right]$

$\Leftrightarrow \varphi\tau_i^f \geq \frac{(1-\gamma)(2\bar{a}-a_i)+\varphi}{(1-\gamma)(2\bar{a}-a_m)+\varphi}\left[(1-\gamma\tau^e)(\bar{a}-a_m) + \varphi\tau_m^f\right] - (1-\gamma\tau^e)(\bar{a}-a_i)$

$\Leftrightarrow \tau_i^f \geq \frac{\frac{(1-\gamma)(2\bar{a}-a_i)+\varphi}{(1-\gamma)(2\bar{a}-a_m)+\varphi}\left[(1-\gamma\tau^e)(\bar{a}-a_m)+\varphi\tau_m^f\right]-(1-\gamma\tau^e)(\bar{a}-a_i)}{\varphi}$

Assume now that a_i and τ_i^f are symmetrically distributed. It follows that the former condition becomes:

$\tau_i^f \geq \frac{\frac{(1-\gamma)(2a_m-a_i)+\varphi}{(1-\gamma)a_m+\varphi}\varphi\tau_m^f-(1-\gamma\tau^e)(a_m-a_i)}{\varphi}$

$\Leftrightarrow \varphi\tau_i^f \geq \left(1 + \frac{(1-\gamma)(a_m-a_i)}{(1-\gamma)a_m+\varphi}\right)\varphi\tau_m^f - (1-\gamma\tau^e)(a_m-a_i)$

$\Leftrightarrow \varphi\left(\tau_i^f - \tau_m^f\right) \geq \frac{(1-\gamma)(a_m-a_i)}{(1-\gamma)a_m+\varphi}\varphi\tau_m^f - (1-\gamma\tau^e)(a_m-a_i)$

$\Leftrightarrow \varphi\left(\tau_i^f - \tau_m^f\right) \geq \left[(1-\gamma\tau^e) - \frac{(1-\gamma)}{(1-\gamma)a_m+\varphi}\varphi\tau_m^f\right](a_i-a_m)$

$\Leftrightarrow \tau_i^f - \tau_m^f \geq \left[\frac{1-\tau_m-\gamma(\tau^e-\tau_m)}{\varphi}\right](a_i-a_m)$

Define $f(x)$ the density function of $x = a - a_m$, $g(z)$ the density function of $z = \tau^f - \tau_m^f$ and $k = \frac{1-\tau_m-\gamma(\tau^e-\tau_m)}{\varphi}$. Both distributions are by assumption symmetrical and centered. It follows that:

$\Pr\{\mathcal{C}\} = \int_{-\infty}^{+\infty}\left[\int_{k(a_i-a_m)}^{+\infty} g(z)\,dz\right] f(x)\,dx$

$= \int_{-\infty}^{0}\left[\int_{kx}^{+\infty} g(z)\,dz\right] f(x)\,dx + \int_{0}^{+\infty}\left[\int_{kx}^{+\infty} g(z)\,dz\right] f(x)\,dx$

$= \int_{-\infty}^{0}[1-G(kx)]\,f(x)\,dx + \int_{0}^{+\infty}[1-G(kx)]\,f(x)\,dx$

where G is the pdf of g.

Let $y = -x$. Symmetry of f and g yields $f(x) = f(-y) = f(y)$ and $1 - G(kx) = 1 - G(-ky) = G(ky)$. It follows that:

$\Pr\{\mathcal{C}\} = \int_{0}^{+\infty} G(ky)\,f(y)\,dy + \int_{0}^{+\infty}[1-G(kx)]\,f(x)\,dx = \int_{0}^{+\infty} f(x)\,dx = \frac{1}{2}$.

a_i and τ_i^f both symmetrically distributed are then sufficient conditions so that the median voter is the condorcet winner.

ABOUT OFCE

The Paris-based Observatoire français des conjonctures économiques (OFCE), or French Economic Observatory is an independent and publicly-funded centre whose activities focus on economic research, forecasting and the evaluation of public policy.

Its 1981 founding charter established it as part of the French Fondation nationale des sciences politiques (Sciences Po), and gave it the mission is to "ensure that the fruits of scientific rigour and academic independence serve the public debate about the economy". The OFCE fulfils this mission by conducting theoretical and empirical studies, taking part in international scientific networks, and assuring a regular presence in the media through close cooperation with the French and European public authorities. The work of the OFCE covers most fields of economic analysis, from macroeconomics, growth, social welfare programmes, taxation and employment policy to sustainable development, competition, innovation and regulatory affairs.

ABOUT SCIENCES PO

Sciences Po is an institution of higher education and research in the humanities and social sciences. Its work in law, economics, history, political science and sociology is pursued through ten research units and several crosscutting programmes.
Its research community includes over two hundred twenty members and three hundred fifty PhD candidates. Recognized internationally, their work covers a wide range of topics including education, democracies, urban development, globalization and public health.
One of Sciences Po's key objectives is to make a significant contribution to methodological, epistemological and theoretical advances in the humanities and social sciences. Sciences Po's mission is also to share the results of its research with the international research community, students, and more broadly, society as a whole.

PARTNERSHIP