American Economic Review 2024, 114(11): 3551–3577
https://doi.org/10.1257/aer.20211025

Generalized Social Marginal Welfare Weights Imply Inconsistent Comparisons of Tax Policies[†]

By Itai Sher*

This paper concerns Saez and Stantcheva's (*2016*) *generalized social marginal welfare weights*, *which aggregate losses and gains due to tax policies while incorporating nonutilitarian ethical considerations. The approach evaluates local tax changes without a global social objective. I show that local tax policy comparisons implicitly entail global comparisons. Moreover*, *whenever welfare weights do not have a utilitarian structure*, *these implied global comparisons are inconsistent. I argue that broader ethical values cannot in general be represented simply by modifying the weights placed on benefits to different people*, *and a more thoroughgoing modification of the utilitarian approach is required.* (*JEL* D60, D63, D71, H21, H23, I31)

The traditional optimal tax literature, building on the classic work of Mirrlees (1971), has adopted a broadly utilitarian normative framework. As argued by several recent authors, including Weinzierl (2014, 2017) and Fleurbaey and Maniquet (2018), the omission of other ethical principles that people care about, such as libertarianism, equality of opportunity, and desert, is a serious problem for the classical approach. Saez and Stantcheva (2016) have proposed a general, relatively simple, way of addressing these concerns. They argue that one can modify the optimality conditions of the standard approach so that these can incorporate broader values while maintaining the structure of the standard optimal taxation theory. According to Saez and Stantcheva's (2016) *generalized social marginal welfare weights* (GSMWW) approach, all one has to do is substitute for the standard utilitarian welfare weights—corresponding to the marginal utility of consumption—other welfare weights reflecting broader values. Such generalized welfare weights can effectively be used as a kind of "get out of jail free" card that allows one to ignore normative issues on the assumption that they can be incorporated simply by appropriate selection of welfare

* Department of Economics, University of Massachusetts Amherst (email: isher@umass.edu). Mikhail Golosov was the coeditor for this article. I am grateful for helpful comments from and discussions with Matthew Adler, Eduardo Davila, Pawel Doligalski, Piotr Dworczak, Maya Eden, Wojciech Kopczuk, Benjamin Lockwood, Juan Moreno-Cruz, Louis Perrault, Paolo Piacquadio, Peter Sher, and Matthew Weinzierl and to seminar audiences at UC Riverside, the Welfare Economics and Economic Policy virtual seminar, the University of Chicago Harris Public Policy's Political Economy Workshop, the PPE Society annual meeting, the Global Priorities Institute at Oxford University, Northwestern University, and King's College London. I would also like to thank the editor and three anonymous referees.

[†] Go to https://doi.org/10.1257/aer.20211025 to visit the article page for additional materials and author disclosure statement(s).

weights.[1] In this paper, I show formally that this solution to the problem of incorporating broader values into optimal tax does not work because it leads to inconsistencies. It is not possible, in general, to capture broad ethical principles simply by means of welfare weights. Broadening the normative considerations that bear on taxation will require a more thoroughgoing revision of optimal tax theory.

I now discuss the specific contributions of this paper. The GSMWW approach only claims to make *local comparisons* among tax policies and accordingly to find local optima. Indeed, Saez and Stantcheva (2016, p. 24) write, "In our approach … there is no social welfare objective primitive that the government maximizes." The first contribution of the paper is to show how to collect local comparisons made by generalized welfare weights into implied global social comparisons. In particular, for any system of welfare weights g, I define social preference and indifference relations $\prec^g$ and $\sim^g$ over tax policies, which capture some of the global social comparisons implied by welfare weights (see Section II). Second, I define a critical property of welfare weights, *structural utilitarianism* (see Section III), which is essential to the question of whether welfare weights are consistent. Third, I show that if welfare weights g are structurally utilitarian, then welfare weights are consistent in the sense that there exists a social welfare function that generates those welfare weights (see Theorem 1 in Section III). More specifically, I show that the case in which welfare weights are structurally utilitarian is the case in which they can be generated by a generalized utilitarian social welfare function of the form $\int F_i(U_i)di$, where U_i represents agent i's utility and $F_i(U_i)$ is an agent-specific monotonic transformation of this utility.[2] Fourth, I show that, when welfare weights are not structurally utilitarian, they are inconsistent in the sense of the following theorem.

Generalized Welfare Weights Inconsistency Theorem: *If welfare weights g are not structurally utilitarian, then they are inconsistent in the sense there exist tax policies T_0, T_1, T_2, T_3, each of which raises the same revenue, and such that welfare weights imply a social preference cycle of the form* $T_0 \prec^g T_1 \sim^g T_2 \prec^g T_3 \sim^g T_0$.

This is Theorem 3. Theorem 2 is a simpler version of the result with a more accessible proof. Putting together the third and fourth contributions, it follows that structural utilitarianism is necessary and sufficient for welfare weights to be consistent. So the generalized welfare weights approach does not meaningfully add anything beyond what is already available by means of a generalized utilitarian social welfare function, a framework that is long established;[3] the additional possibilities offered by generalized welfare weights are inconsistent.

Some ethical values can be captured in a generalized utilitarian framework by making the transformations F_i suitably dependent on agent characteristics. But Saez and Stantcheva (2016) suggest that libertarian values can be captured by

[1] I am grateful to an anonymous referee for suggesting the formulation of the problem in this paragraph as well as some of the wording.

[2] For utilitarianism to be meaningful and for generalized utilitarianism to be meaningfully different than utilitarianism, we must assume that we are given utilities U_i that are cardinal and interpersonally comparable.

[3] Indeed, Mirrlees (1971) posited a generalized utilitarian social welfare function, although with a common transformation $F(U_i)$ of utility for all agents i.

making welfare weights a function of total taxes paid or that a poverty alleviation imperative can be captured by making weights a function of consumption, and I show that such weights lead to inconsistent judgments (see Sections IVB and VD). Section VI explains how my analysis generalizes when the assumption of quasi-linear preferences, maintained through most of the paper, is dropped. Section VII continues the discussion of the significance of these results and their relation to the literature.

I. Model

This section presents the model of Saez and Stantcheva (2016). I assume all functions are *smooth*—meaning infinitely differentiable—unless their domain is discrete or explicitly stated otherwise.

A. *Standard Aspects of the Model*

There is a continuum of agents uniformly distributed on the interval $I = [0,1]$. Each agent $i \in I$ has *observable characteristics* x_i drawn from the set X and *unobservable characteristics* y_i drawn from the set Y. I assume X and Y are either discrete or subsets of Euclidean spaces. Let c_i be agent i's consumption and z_i be agent i's income. Consumption belongs to the real line $\mathbb{R}$, and income to the nonnegative reals $Z = \mathbb{R}_+$. Agent i has the quasi-linear utility function $U_i(c_i, z_i) = u(c_i - v_i(z_i))$, where $v_i(z_i) = v(z_i, x_i, y_i)$ is the cost of earning income z_i given characteristics (x_i, y_i). Assume $v_i'(z_i) > 0, v_i''(z_i) > 0$, for all z_i, so that v_i is increasing and strictly convex in z_i, $v_i'(z_i) > 1$ for sufficiently large z_i, and that u is increasing and strictly concave. Assume for simplicity that, for all i, $v_i'(0) < 1$, so that in the absence of taxes, all agents earn a positive income. A *tax policy* is a function $T : Z \times X \to \mathbb{R}$, where $T(z,x)$ is the tax paid by agents with income z given observable characteristics x. I write $T_i(z_i) = T(z_i, x_i)$ so that T_i gives i's personalized tax on the basis of i's observable characteristics. I assume tax policies have the formal structure requisite to support the exposition that follows. Section IIA makes more precise assumptions about the set of tax policies for my formal results. Given a tax policy T, we have $c_i = z_i - T_i(z_i)$. Define $z_i(T)$ to be i's optimal income when facing tax policy T, and $c_i(T) = z_i(T) - T_i(z_i(T))$; formally, $z_i(T) \in \operatorname{argmax}_{z_i} U_i(z_i - T_i(z_i), z_i)$. The agent's indirect utility from tax policy T is then $U_i(T) = U_i(c_i(T), z_i(T))$. Let $R(T) = \int T_i(z_i(T))di$ be the revenue generated by T.

B. *Generalized Welfare Weights*

The novelty in the GSMWW approach is the way that tax systems are evaluated. We assume a system $g(c_i, z_i; x_i, y_i)$ of *generalized social marginal welfare weights*. Thus, we assign a certain weight to each agent depending on their consumption c_i, income z_i, and characteristics x_i, y_i. Formally, a system of generalized social welfare weights is a function $g : \mathbb{R} \times Z \times X \times Y \to \mathbb{R}$ such that $g(c_i, z_i; x_i, y_i) > 0$, $\forall c_i, z_i, x_i, y_i$. Define $g_i(c_i, z_i) = g(c_i, z_i; x_i, y_i)$. The intuitive interpretation of generalized social marginal welfare weights is that they measure the marginal social

value of consumption for each person i, and ratios of welfare weights $g_i(c_i, z_i)/g_j(c_j, z_j)$ measure social marginal rates of substitution of consumption for agents i and j. Given a tax system T, the local marginal welfare weight $g_i(T) = g_i\big(c_i(T), z_i(T)\big)$ is endogenously determined. The key innovation of the approach is to assess small tax reforms via local marginal welfare weights rather than by reference to a global objective.

I now present some illustrative examples from Saez and Stantcheva (2016). *Utilitarian weights*: $g_i(c_i, z_i) = \frac{\partial}{\partial c_i} U_i(c_i, z_i) = u'\big(c_i - v_i(z_i)\big)$. These are the standard utilitarian weights that prioritize benefits according to the marginal utility of consumption. *Libertarian weights:* $g_i(c_i, z_i) = \hat{g}(z_i - c_i) = \hat{g}(t_i)$, where $t_i = z_i - c_i$ is the tax paid and we assume that $\hat{g}'(t_i) > 0$. That is, the more tax a person has already paid, the greater the weight placed on that person. *Libertarian-utilitarian mix*: $g_i(c_i, z_i) = \hat{g}\big(c_i - v_i(z_i), z_i - c_i\big) = \hat{g}(\hat{u}_i, t_i)$, where $\hat{u}_i = c_i - v_i(z_i)$ with $\partial\hat{g}/\partial\hat{u}_i < 0$ and $\partial\hat{g}/\partial t_i > 0$. The first inequality can be interpreted as saying weights are increasing in marginal utility for consumption (since $u'\big(c_i - v_i(z_i)\big)$ is decreasing in $c_i - v_i(z_i)$), and the second says that they are also increasing in taxes paid. *Poverty alleviation*: $g(c_i, z_i) = 1$ if $c_i < \bar{c}$, where $\bar{c}$ is the poverty threshold and $g(c_i, z_i) = 0$ otherwise; that is, we put positive and equal weight on those beneath the poverty line, and no weight on those above the poverty line.[4] *Counterfactuals*: Welfare weights can be made to depend on how much someone would have worked in the absence of taxes (which depends on their type) in comparison to how much they work in the presence of taxes. *Equality of opportunity*: Weights can be made to depend on one's rank in the income distribution conditional on one's background conditions. Such weights go beyond the formal framework in that they depend on the entire income distribution and not just on c_i, z_i, x_i, and y_i; Saez and Stantcheva (2016) present several examples that go beyond the basic formal framework they present.

C. *Local Optimality and Local Improvements*

A *tax reform* is a function $\Delta T : Z \times X \to \mathbb{R}$, satisfying appropriate regularity conditions,[5] whose interpretation is that it represents some change to the status quo tax policy. Define $\Delta T_i(z_i) = \Delta T(z_i, x_i)$. Say tax reform ΔT is *locally budget neutral* at tax policy T if $\frac{d}{d\varepsilon}\Big|_{\varepsilon=0} R(T + \varepsilon\Delta T) = 0$. Say that a locally budget-neutral tax reform ΔT is *locally desirable* at T if

$$\int g_i(T)\Delta T_i\big(z_i(T)\big)di < 0. \tag{1}$$

In other words, ΔT is locally desirable at T if the cost of the tax change to different individuals due to a small version of the reform $\varepsilon\Delta T$, weighted by the local welfare weights, is negative. Say that tax system T satisfies the *local optimal tax criterion*

[4] Such weights are only assumed to be nonnegative but not positive everywhere, contrary to our assumption.

[5] See Section IIA, which imposes formal regularity conditions on parameterized families of tax policies (T^θ); tax policies modified by small reforms $T + \theta\Delta T$ are a special case of parameterized tax policies T^θ.

if, for all locally budget-neutral tax reforms ΔT, $\int g_i(T)\Delta T_i(z_i(T))di = 0$. Saez and Stantcheva (2016) say that this criterion gives a necessary condition for local optimality of a tax system T and use it to derive optimal tax formulas for generalized welfare weights that are analogous to the standard optimal tax formulas.

In the traditional utilitarian framework, the goal is to choose a tax policy T to maximize the utilitarian objective $\int U_i(T)di$ subject to a revenue requirement. Given this formulation, employing utilitarian weights $g_i(c_i, z_i) = \frac{\partial}{\partial c_i} U_i(c_i, z_i)$, and using the envelope theorem, the local optimal tax criterion is a necessary condition for T to be an optimum, and (1) is a sufficient condition for a small version of the reform ΔT to be a local improvement. However, in the GSMWW framework, there is no global objective from which to derive these conditions; so the conditions for a locally desirable reform and for a local optimum are posited by analogy to the utilitarian case.

II. Global Social Comparisons Implied by Welfare Weights

Generalized social marginal welfare weights provide *local* comparisons: conditions for a local improvement and for local optimality of tax policies. This section shows how to derive *global* social comparisons implicit in welfare weights.

A. *Modifying Tax Policies*

To derive global comparisons, I need to smoothly vary tax policies in a parametric way. To do so, I append a (real-valued) parameter θ to tax policies, writing T^θ. Varying θ corresponds to changing tax policy in some way. For example, if $T^\theta = T + \theta\Delta T$, then θ measures the size of the tax reform ΔT to tax policy T. Alternatively, consider a (nonindividualized) linear tax $T^\theta(z) = \theta z + \kappa(\theta)$, where, when we vary θ, we vary both the marginal tax rate and the lump-sum tax $\kappa(\theta)$. In general, let $\Theta = \left[\underline{\theta}, \bar{\theta}\right]$ be an interval in the real line, where $\underline{\theta} < \bar{\theta}$. Consider a parameterized collection $\left(T^\theta\right)_{\theta\in\Theta}$ of tax policies. Below, I sometimes use the abbreviated notation $\left(T^\theta\right)$ rather than $\left(T^\theta\right)_{\theta\in\Theta}$. Given $\left(T^\theta\right)$, define $T_i(z,\theta) = T_i^\theta(z)$ so that $T_i(z,\theta)$ can be regarded as a real-valued function with domain $Z \times \Theta$. A family of tax policies $\left(T^\theta\right)_{\theta\in\Theta}$ is *well behaved* if (i) for each i and θ, i's optimal income in response to T^θ exists, is unique, and positive, and the second-order condition for i's optimization problem, when facing T^θ, holds with strict inequality at the optimum, and (ii) for all i, the map $(z,\theta) \mapsto T_i(z,\theta)$ is smooth, and, except for at most at finitely many values of i, the map $(i,z,\theta) \mapsto T_i(z,\theta)$ is smooth. Note that the second condition allows that, when taxes are individualized, there may be finitely many i such that tax policy is discontinuous at i. Say a tax policy T is *regular* if there exists a well-behaved family $\left(T^\theta\right)_{\theta\in\Theta}$ and $\theta' \in \Theta$ such that $T^{\theta'} = T$. Regular tax policies are characterized by conditions similar to (i) and (ii) above (see online Appendix A.1). Given a family $\left(T^\theta\right)$, write $z_i(\theta) = z_i\left(T^\theta\right), c_i(\theta) = c_i\left(T^\theta\right)$, $U_i(\theta) = U_i\left(T^\theta\right), g_i(\theta) = g_i\left(T^\theta\right)$ for, respectively, i's optimal income, optimal consumption, indirect utility, and welfare weight at T^θ.

Sometimes I introduce a second parameter ϵ in $E = \left[\underline{\epsilon}, \bar{\epsilon}\right]$, where $\underline{\epsilon} < \bar{\epsilon}$, and consider a doubly parameterized family $\left(T^{\theta,\epsilon}\right)_{\theta\in\Theta,\epsilon\in E}$ (abbreviated as $\left(T^{\theta,\epsilon}\right)$). As above, I write $T_i(z,\theta,\epsilon) = T_i^{\theta,\epsilon}(z)$. $\left(T^{\theta,\epsilon}\right)$ is *well behaved* if it satisfies conditions

analogous to (i) and (ii) above, with (θ, ϵ) playing the role of θ, so that, for example, the first part of (ii) becomes for all i, the map $(z, \theta, \epsilon) \mapsto T_i(z, \theta, \epsilon)$ is smooth. For a complete definition, see online Appendix A.1. Given family $(T^{\theta,\epsilon})$, write $z_i(\theta, \epsilon)$, $c_i(\theta, \epsilon), U_i(\theta, \epsilon)$, and $g_i(\theta, \epsilon)$ for i's optimal income, optimal consumption, indirect utility, and welfare weight at $T^{\theta,\epsilon}$.

B. *The Global Improvement and Indifference Principles*

Consider a system of generalized social welfare weights g. I now define a relation $\prec^g$, which captures some of the strict social preferences implied by g, and a relation $\sim^g$, which captures some of the social indifferences implied by g.[6] For any pair of tax policies T_0 and T_1, whenever $T_0 \prec^g T_1$, this indicates that welfare weights g imply that T_1 is strictly socially preferred to T_0, and whenever $T_0 \sim^g T_1$, this indicates that welfare weights g imply that T_1 is socially indifferent to T_0.

Let $(T^\theta)_{\theta \in \Theta}$ be a well-behaved parameterized collection of tax policies, and let $\theta_0, \theta_1 \in \Theta$ be such that $\theta_0 < \theta_1$. Consider the following principles.

Global Improvement Principle: Suppose that, for all $\hat{\theta} \in [\theta_0, \theta_1]$,

$$\int g_i(\hat{\theta}) \frac{\partial}{\partial \theta}\Big|_{\theta=\hat{\theta}} T_i\big(z_i(\hat{\theta}), \theta\big) di < 0. \tag{2}$$

That is, increasing θ is locally socially desirable at $\hat{\theta}$. Then $T^{\theta_0} \prec^g T^{\theta_1} : T^{\theta_1}$ is socially preferred to T^{θ_0}.

Global Indifference Principle: Suppose that, for all $\hat{\theta} \in [\theta_0, \theta_1]$,

$$\int g_i(\hat{\theta}) \frac{\partial}{\partial \theta}\Big|_{\theta=\hat{\theta}} T_i\big(z_i(\hat{\theta}), \theta\big) di = 0. \tag{3}$$

That is, at $\hat{\theta}$, welfare weights don't detect any change in social welfare as θ changes. Then $T^{\theta_0} \sim^g T^{\theta_1}$: T^{θ_0} and T^{θ_1} are socially indifferent.

We can think of these two principles as axioms that allow us to draw inferences about social preferences from welfare weights. Henceforth, I shall assume that $\prec^g$ and $\sim^g$ satisfy these principles.

To understand these principles, consider first the standard utilitarian case. The utilitarian social welfare of tax policy T^θ is $W_{\text{util}}(\theta) = \int U_i(\theta) di$. Because

$$U_i(\theta) = U_i\Bigg(\underbrace{z_i(\theta) - T_i\big(z_i(\theta), \theta\big)}_{c_i(\theta)}, z_i(\theta)\Bigg),$$

[6] I do not claim that $\prec^g$ and $\sim^g$ capture *all* social preferences implicit in welfare weights g.

it follows from the envelope theorem that, for any $\hat{\theta} \in \Theta$,

$$\frac{d}{d\theta} U_i\big(\hat{\theta}\big) = \underbrace{-\frac{\partial}{\partial c_i} U_i\Big(c_i\big(\hat{\theta}\big), z_i\big(\hat{\theta}\big)\Big)}_{\text{utilitarian welfare weight}} \underbrace{\frac{\partial}{\partial \theta}\Big|_{\theta=\hat{\theta}} T_i\Big(z_i\big(\hat{\theta}\big), \theta\Big)}_{\text{direct effect on taxes}}. \tag{4}$$

That is, the envelope theorem tells us that the marginal effect of a change in tax policy on an agent's utility is the product of the agent's marginal utility of consumption and the marginal direct effect of the change in θ on the agent's tax bill, and we can ignore the indirect effects due to changes in behavior (the choices of consumption and income) as taxes change. So in the utilitarian case,

$$\begin{aligned}
\frac{d}{d\theta} W_{\text{util}}\big(\hat{\theta}\big) &= \int \frac{d}{d\theta} U_i\big(\hat{\theta}\big) di \\
&= -\int \frac{\partial}{\partial c_i} U_i\Big(c_i\big(\hat{\theta}\big), z_i\big(\hat{\theta}\big)\Big) \frac{\partial}{\partial \theta}\Big|_{\theta=\hat{\theta}} T_i\Big(z_i\big(\hat{\theta}\big), \theta\Big) di \\
&= -\int g_i\big(\hat{\theta}\big) \frac{\partial}{\partial \theta}\Big|_{\theta=\hat{\theta}} T_i\Big(z_i\big(\hat{\theta}\big), \theta\Big) di.
\end{aligned}$$

Given this equation, (2) becomes $\frac{d}{d\theta} W_{\text{util}}\big(\hat{\theta}\big) > 0$, and (3) becomes $\frac{d}{d\theta} W_{\text{util}}\big(\hat{\theta}\big) = 0$, so that the global improvement principle says that if utilitarian welfare is increasing as we vary θ from θ_0 to θ_1, then utilitarian welfare is greater at θ_1 than at θ_0, and the global indifference principle says that if utilitarian welfare is unchanging as we vary θ, then utilitarian welfare is the same at θ_1 as at θ_0. In the utilitarian case, these principles are obviously valid.

In the case of generalized welfare weights, the global improvement and indifference principles are posited by analogy with the utilitarian case. This is the same as the justification for Saez and Stantcheva's (2016) definitions for a local desirability of a tax reform and local optimality of a tax policy, which substitute generalized welfare weights $g_i\big(\hat{\theta}\big)$ for utilitarian welfare weights $\frac{\partial}{\partial c_i} U_i\Big(c_i\big(\hat{\theta}\big), z_i\big(\hat{\theta}\big)\Big)$ in principles that are valid for utilitarianism. Indeed, when the parameterized family of tax policies has the form $T^\theta = T + \theta \Delta T$ and $\hat{\theta} = 0$, (2) simplifies to (1) in Section IC, Saez and Stantcheva's (2016) condition for a locally desirable tax reform.[7]

The following useful result assumes the global improvement principle and follows from our smoothness assumptions; see the online Appendix for the proof.

PROPOSITION 1 (Local Improvement Principle): *Let g be a system of welfare weights, let $\big(T^\theta\big)_{\theta \in [\underline{\theta}, \bar{\theta}]}$ be a well-behaved parameterized family of tax policies, and let $\theta_0 \in [\underline{\theta}, \bar{\theta})$. If $\int g_i(\theta_0) \frac{\partial}{\partial \theta}\Big|_{\theta=\theta_0} T_i\big(z_i(\theta_0), \theta\big) di < 0$, then there exists $\theta_1 \in \big(\theta_0, \bar{\theta}\big]$ such that, for all $\theta \in (\theta_0, \theta_1)$, $T^{\theta_0} \prec^g T^\theta$. Similarly, if $\int g_i(\theta_0) \frac{\partial}{\partial \theta}\Big|_{\theta=\theta_0} T_i\big(z_i(\theta_0), \theta\big) di > 0$, then there exists $\theta_1 \in \big(\theta_0, \bar{\theta}\big]$ such that, for all $\theta \in (\theta_0, \theta_1)$, $T^{\theta_0} \succ^g T^\theta$.*

[7] Saez and Stantcheva (2016) apply this condition to locally revenue-neutral tax reforms; in my main theorem, I use the global improvement and indifference principles to construct a cycle when revenue remains constant.

C. *Pareto*

Certain Pareto conditions, which are useful below, are implicit in the welfare weights framework. In particular, it follows from (4), which was derived using the envelope theorem, and the fact that the marginal utility of consumption is positive that the following relation holds:

$$\forall i,\ \forall \hat{\theta}, \frac{d}{d\theta} U_i\left(\hat{\theta}\right) \gtreqless 0 \Leftrightarrow \frac{\partial}{\partial \theta}\Big|_{\theta=\hat{\theta}} T_i\left(z_i\left(\hat{\theta}\right), \theta\right) \gtreqless 0. \tag{5}$$

That is, $\frac{d}{d\theta} U_i\left(\hat{\theta}\right)$ and $\frac{\partial}{\partial \theta}\Big|_{\theta=\hat{\theta}} T_i\left(z_i\left(\hat{\theta}\right), \theta\right)$ always have the opposite sign when non-zero, and otherwise both are zero. This shows that the term $-\frac{\partial}{\partial \theta}\Big|_{\theta=\hat{\theta}} T_i\left(z_i\left(\hat{\theta}\right), \theta\right)$ captures preferences in the sense that it points in the same direction as preferences do in response to a change in θ; and it also shows why the global improvement and indifference principles respect preferences. If, at $\hat{\theta}$, an increase in θ makes all agents better-off, the terms $\frac{\partial}{\partial \theta}\Big|_{\theta=\hat{\theta}} T_i\left(z_i\left(\hat{\theta}\right), \theta\right)$ will be negative for all agents, and so $\int g_i\left(\hat{\theta}\right) \frac{\partial}{\partial \theta}\Big|_{\theta=\hat{\theta}} T_i\left(z_i\left(\hat{\theta}\right), \theta\right) di < 0$, no matter what (positive) welfare weights g are used. This is formalized by the following proposition, which is proved in the online Appendix and assumes, as above, that welfare weights are always positive and also assumes the global improvement and indifference principles.

PROPOSITION 2: *Let* $\left(T^{\theta}\right)_{\theta \in \Theta}$ *be a well-behaved family of tax policies, and let* $\theta_0, \theta_1 \in \Theta$ *be such that* $\theta_0 < \theta_1$.

(*i*) **Pareto Indifference along Paths:** *Suppose that all agents are indifferent among all tax policies* T^{θ} *for* $\theta \in [\theta_0, \theta_1]$. *Then, for all systems of welfare weights* g, $T^{\theta_0} \sim^{g} T^{\theta_1}$.

(*ii*) **Weak Pareto along Paths:** *Suppose that for all* $\hat{\theta} \in [\theta_0, \theta_1]$ *and all agents* i, $\frac{d}{d\theta} U_i\left(\hat{\theta}\right) > 0$ *so that, for all agents, tax policies become more preferred as* θ *increases within* $[\theta_0, \theta_1]$. *Then, for all systems of welfare weights* g, $T^{\theta_0} \prec^{g} T^{\theta_1}$.

The Pareto principles stated above are weaker than the standard principles because they only apply to paths of smoothly changing tax policies along which the direction of preferences is constant. Say that a social welfare function is *Paretian along paths* if it satisfies a weakened version of the Pareto principle, analogous to the properties that the above proposition shows to be satisfied by all systems of welfare weights. A formal statement of this property of social welfare functions, as well as of what it means for a system of welfare weights to implement a social welfare function and a proof of the following corollary, is in the online Appendix.

COROLLARY 1: *Any social welfare function that is not Paretian along paths cannot be implemented by any system of generalized social welfare weights.*

The corollary shows that the expressive power of welfare weights is limited in the sense that non-Paretian (in a weak sense of Paretian) objectives cannot be implemented by welfare weights.

III. Structural Utilitarianism

The key condition for generalized welfare weights to be consistent is *structural utilitarianism*.

DEFINITION 1: *A system of welfare weights g is* ***structurally utilitarian*** *if and only if* $\forall i \in I,\ \forall z_i,\ z_i' \in Z,\ \forall c_i, c_i' \in \mathbb{R}$,

$$c_i - v_i(z_i) = c_i' - v_i(z_i') \Rightarrow g_i(c_i, z_i) = g_i(c_i', z_i'). \tag{6}$$

To interpret this definition, observe that, given quasi-linear utility $U_i(c_i, z_i) = u\big(c_i - v_i(z_i)\big)$, we have $\frac{\partial}{\partial c_i} U_i(c_i, z_i) = u'\big(c_i - v_i(z_i)\big)$. Thus, the marginal utility of consumption $\frac{\partial}{\partial c_i} U_i(c_i, z_i)$ is determined by the quantity $c_i - v_i(z_i)$, and given our assumption that the outer utility function $u(\cdot)$ is strictly concave, the condition (6) for structural utilitarianism is equivalent to

$$\frac{\partial}{\partial c_i} U_i(c_i, z_i) = \frac{\partial}{\partial c_i} U_i(c_i', z_i') \Rightarrow g_i(c_i, z_i) = g_i(c_i', z_i'). \tag{7}$$

Thus, structural utilitarianism allows that welfare weights are not necessarily *equal to* the marginal utility of consumption $\frac{\partial}{\partial c_i} U_i(c_i, z_i)$, the utilitarian welfare weight, but it requires that welfare weights are *determined by* the marginal utility of consumption in the sense that, if i's marginal utility of consumption does not change, then i's welfare weight does not change. Note that the condition is imposed separately on each agent i; it is a condition on how that agent's welfare weight changes as their allocation (c_i, z_i) changes, and no relation is posited between the welfare weights of different agents i and j. So structural utilitarianism is consistent with welfare weights being dependent on agents' characteristics (x_i, y_i). Recalling that utility is given by $U_i(c_i, z_i) = u\big(c_i - v_i(z_i)\big)$, utility is also determined by the quantity $c_i - v_i(z_i)$. When the outer utility function $u(\cdot)$ is both strictly increasing and strictly concave, $\frac{\partial}{\partial c_i} U_i(c_i, z_i) = \frac{\partial}{\partial c_i} U_i(c_i', z_i')$ if and only if $U_i(c_i, z_i) = U_i(c_i', z_i')$, and the condition (6) for structural utilitarianism is also equivalent to

$$U_i(c_i, z_i) = U_i(c_i', z_i') \Rightarrow g_i(c_i, z_i) = g_i(c_i', z_i'). \tag{8}$$

Thus, structural utilitarianism can also be interpreted as saying that i's welfare weight doesn't change when i's utility doesn't change. The coincidence of (7) and (8) depends on the assumption of quasi-linear utility, and, indeed, in Section VI, I show how to generalize structural utilitarianism when utility is no longer assumed quasi-linear.

Define $\hat{U}_i(c_i, z_i) = c_i - v_i(z_i)$. $\hat{U}_i(c_i, z_i)$ is a utility function over (c_i, z_i) pairs that is ordinally equivalent to $U_i(c_i, z_i)$. Define the variable $\hat{u}_i$ by $\hat{u}_i = \hat{U}_i(c_i, z_i)$. We can then reexpress welfare weights as a function $\hat{g}_i(\hat{u}_i, z_i)$ of utility and income $(\hat{u}_i, z_i)$ rather than as a function $g_i(c_i, z_i)$ of consumption and income (c_i, z_i). The relationship between the two expressions is as follows:

$$\hat{g}_i(\hat{u}_i, z_i) = g_i\big(\hat{u}_i + v_i(z_i), z_i\big), \ \forall \hat{u}_i \in \mathbb{R}, \ \forall z_i \in Z. \tag{9}$$

The following result is useful. (The straightforward proof is in the online Appendix.)

PROPOSITION 3: *Let g and $\hat{g}$ be related as in* (9). *Then welfare weights g are structurally utilitarian if and only if* $\forall i \in I, \ \forall \hat{u}_i \in \mathbb{R}, \ \forall z_i \in Z, \ \frac{\partial}{\partial z_i}\hat{g}_i(\hat{u}_i, z_i) = 0$.

We now come to a theorem that shows when welfare weights are structurally utilitarian, they correspond to a global social ranking. Say that a real-valued function $W(T)$, whose domain is the set of regular tax policies, is a *generalized utilitarian social welfare function* if there exists a real-valued function $F_i(u_i) = F(u_i, x_i, y_i)$, which is (i) smooth in u_i and smooth in (u_i, x_i, y_i) unless (x_i, y_i) are discrete and (ii) strictly increasing in u_i, such that $W(T) = \int F_i\big(U_i(c_i(T), z_i(T))\big)di$.[8] It follows from the envelope theorem that, for all well-behaved families $(T^\theta)_{\theta\in\Theta}$ and $\theta_0 \in \Theta$, $\frac{d}{d\theta}\big|_{\theta=\theta_0} W(T^\theta) = -\int F_i'\big(U_i(c_i(\theta_0), z_i(\theta_0))\big)\frac{\partial}{\partial c_i}U_i(c_i(\theta_0), z_i(\theta_0))\frac{\partial}{\partial \theta}\big|_{\theta=\theta_0} T_i(z_i(\theta_0), \theta)di$. Hence, $F_i'\big(U_i(c_i, z_i)\big)\frac{\partial}{\partial c_i}U_i(c_i, z_i)$ are the social welfare weights arising from a generalized utilitarian social welfare function. Formally, say that a system of welfare weights *g arise from a generalized utilitarian social welfare function* if there exists $F_i(u_i) = F(u_i, x_i, y_i)$ satisfying properties (i) and (ii) above such that for all i, c_i, and z_i, $g_i(c_i, z_i) = F_i'\big(U_i(c_i, z_i)\big)\frac{\partial}{\partial c_i}U_i(c_i, z_i)$.

THEOREM 1: *Welfare weights g are structurally utilitarian if and only if they arise from a generalized utilitarian social welfare function.*[9]

The theorem has the following important corollary.

COROLLARY 2: *If welfare weights g are structurally utilitarian, then there exists a generalized utilitarian social welfare function W from which the welfare weights can be derived in the sense that for all well-behaved families $(T^\theta)_{\theta\in\Theta}$ and $\theta_0 \in \Theta$, $\frac{d}{d\theta}\big|_{\theta=\theta_0} W(T^\theta) = -\int g_i(T^{\theta_0})\frac{\partial}{\partial\theta}\big|_{\theta=\theta_0} T_i\big(z_i(T^{\theta_0}), \theta\big)di$, so that the welfare weights correspond to a consistent social ranking.*

[8] Note that I build smoothness into the definition of a generalized utilitarian social welfare function because I assumed similar smoothness properties on welfare weights. If the smoothness requirements on welfare weights were relaxed somewhat, one could correspondingly weaken the smoothness requirements for a generalized utilitarian social welfare function and still prove a corresponding version of Theorem 1.

[9] One might wonder why the social welfare function in the above theorem is additively separable; the answer is that i's welfare weight is assumed to depend only on i's consumption, income, and characteristics, and not on the distribution of these in society.

Both the theorem and the corollary are proved in the online Appendix. It should be clear that if welfare weights arise from a social welfare function, then it is not possible to use them to construct a social preference cycle. A proof sketch of Theorem 1 is as follows. First, if welfare weights arise from a generalized utilitarian social welfare function, then they are of the form $g_i(c_i,z_i) = F_i'\big(U_i(c_i,z_i)\big)\frac{\partial}{\partial c_i}U_i(c_i,z_i)$. These weights are structurally utilitarian because, given quasi-linearity, both $U_i(c_i,z_i)$ and $\frac{\partial}{\partial c_i}U_i(c_i,z_i)$ are determined by $c_i - v_i(z_i)$. Going in the other direction, by Proposition 3, structural utilitarianism is equivalent to the requirement that welfare weights are a function of $\hat{u}_i = c_i - v_i(z_i)$, so that, assuming structural utilitarianism, we can write $g_i(c_i,z_i) = \hat{g}_i\big(c_i - v_i(z_i)\big) = \hat{g}_i(\hat{u}_i)$. Define the function $w_i(\hat{u}_i)$ by $w_i\big(\hat{u}_i^0\big) = \int_0^{\hat{u}_i^0}\hat{g}_i(\hat{u}_i)\,d\hat{u}_i$. Then define the utility function $W_i(c_i,z_i) = w_i\big(c_i - v_i(z_i)\big)$. Observe that the utility function $W_i(c_i,z_i)$ is ordinally equivalent to $U_i(c_i,z_i)$ in the sense that the two represent the same preferences over consumption and income. It follows that there exists a strictly increasing function F_i such that $W_i(c_i,z_i) = F_i\big(U_i(c_i,z_i)\big)$. Since $g_i(c_i,z_i) = g(c_i,z_i,x_i,y_i)$, there exists some function F such that $F_i(u_i) = F(u_i,x_i,y_i)$. By construction, $g_i(c_i,z_i) = \hat{g}_i\big(c_i - v_i(z_i)\big) = w_i'\big(c_i - v_i(z_i)\big) = \frac{\partial}{\partial c_i}W_i(c_i,z_i) = F_i'\big(U_i(c_i,z_i)\big)\frac{\partial}{\partial c_i}U_i(c_i,z_i)$, which is what we need to show.

IV. A Simple Version of the Main Theorem

A. *The Special Case When Taxes Can Be Completely Individualized*

I now prove a simplified version of my main result. A stronger version is in Section V. Consider the special case in which taxes can be completely individualized so that each agent i faces an individualized tax schedule T_i^θ that can differ from the tax schedule faced by other agents. In our framework, this is possible if each agent's observable characteristics uniquely identify them: formally, for all $i,j \in I, i \neq j \Rightarrow x_i \neq x_j$. I assume that the map $i \mapsto x_i$ is smooth, that there are no unobservable characteristics y_i, and that the functions $u(\cdot)$, $(z_i,x_i) \mapsto v(z_i,x_i)$, $(c_i,z_i,x_i) \mapsto g(c_i,z_i,x_i)$ are smooth. This case is not interesting from an optimal tax perspective because we can simply set the marginal tax rate equal to zero for each agent so that all agents earn the efficient level of income and we can meet the revenue requirement and achieve any redistribution we wish via individualized lump-sum taxes. However, the assumption of completely individualized taxes does allow us to illustrate the problems with welfare weights in a simple way.

THEOREM 2: *Suppose that taxes can be completely individualized. If welfare weights g are not structurally utilitarian, then they are inconsistent in the sense there exist tax policies T_0,T_1,T_2,T_3, each of which raises the same revenue, and such that welfare weights imply a social preference cycle of the form $T_0 \prec^g T_1 \sim^g T_2 \prec^g T_3 \sim^g T_0$.*

A proof sketch follows. Assume welfare weights are not structurally utilitarian. Then it is possible to construct a completely individualized family $\big(T^\theta\big)$ of tax policies such

that for some set S of agents, where both S and the set of agents not in S have positive measure, we have

(i) for agents not in S, taxes are completely unchanged as θ varies;

(ii) for agents in S, the optimal response (c_i, z_i) to taxes changes as θ changes in such a way that the aggregate welfare weight on S, $g_S = \int_S g_i(c_i, z_i)di$, changes but the utility of each agent i is unchanged, so that agents in S are indifferent about the value of θ.

That it is possible to construct a family with the second property follows from the assumption that welfare weights are not structurally utilitarian. The characterization (8) of structural utilitarianism implies that, when welfare weights are not structurally utilitarian, for some agent i, it is possible to vary (c_i, z_i) in such a way that utility $U_i(c_i, z_i)$ does not change but the welfare weight $g_i(c_i, z_i)$ changes. By the smoothness of welfare weights and utility functions, this holds for all agents in a neighborhood S of i, and we may choose the neighborhood so that $g_i(c_i, z_i)$ changes in the same direction for all agents in S as θ changes, and hence, the aggregate welfare weight g_S changes as well. These changes can be brought about as optimal responses to a linear tax individualized policy (for agents in S), $T_i^\theta(z_i) = \tau_i(\theta)z_i + \kappa_i(\theta)$, where the marginal tax rate $\tau_i(\theta)$ controls the choice pretax income z_i and consumption c_i is brought to desired level by the lump-sum tax $\kappa_i(\theta)$. In the above construction, all agents are indifferent as θ changes. So it follows from part (i) of Proposition 2—Pareto indifference along paths—that, letting θ vary from θ_0 to θ_1, welfare weights will imply that the resulting change is socially indifferent:

$$T^{\theta_0} \sim^g T^{\theta_1}. \tag{10}$$

Let O be a positive measure set of agents that is disjoint from S, and such that the set of agents outside of both S and O has positive measure. By our assumptions, the aggregate welfare weight g_S on agents in S changes as θ varies between θ_0 and θ_1, while the aggregate welfare weight $g_O = \int_O g_i(c_i, z_i)di$ on agents in O does not change. It follows that the social marginal rate of substitution g_S/g_O of consumption of agents in S for consumption of agents in O changes as θ moves from θ_0 to θ_1. Assume without loss of generality that g_S increases as θ increases. It follows that there exists some pair of payments t_S and t_O, such that, for sufficiently small $\epsilon > 0$, increasing taxes for agents in S by ϵt_S lump-sum while reducing the taxes of agents in O by ϵt_O lump-sum is desirable at θ_0 and undesirable at θ_1. Formally, define $T^{\theta,\epsilon}$ by

$$T_i^{\theta,\epsilon}(z_i) = \begin{cases} T_i^\theta(z_i) + \epsilon t_S, & \text{if } i \in S; \\ T_i^\theta(z_i) - \epsilon t_O, & \text{if } i \in O; \\ T_i^\theta(z_i), & \text{otherwise.} \end{cases} \tag{11}$$

It then follows that if t_S and t_O are chosen as described above, then for sufficiently small $\epsilon > 0$,

$$T^{\theta_0} \prec^g T^{\theta_0,\epsilon}, \quad T^{\theta_1} \succ^g T^{\theta_1,\epsilon}. \tag{12}$$

Formally, this part of the argument appeals to Proposition 1, the local improvement principle. $T^{\theta,\epsilon}$ differs from T^{θ} for each agent i at most by a change in the lump-sum payment that is independent of θ. Because utility is quasi-linear, $T^{\theta,\epsilon}$ then inherits from T^{θ} the property that each agent is indifferent as θ changes, so that again by Pareto indifference along paths (Proposition 2),

$$T^{\theta_0,\epsilon} \sim^g T^{\theta_1,\epsilon}. \tag{13}$$

Putting (10),(12), and (13) together, we have that for sufficiently small $\epsilon > 0$,

$$T^{\theta_0} \prec^g T^{\theta_0,\epsilon} \sim T^{\theta_1,\epsilon} \prec^g T^{\theta_1} \sim^g T^{\theta_0}. \tag{14}$$

So on the assumption that welfare weights are not structurally utilitarian, we have constructed a social preference cycle.

The last step is to show that revenue can be held fixed across the tax policies in the cycle. This requires a modification of the tax policies T^{θ} and $T^{\theta,\epsilon}$. Observe that $T^{\theta} = T^{\theta,\epsilon}$ when $\epsilon = 0$, so we can identify T^{θ} and $T^{\theta,0}$. Now consider a positive measure set of agents Q, which is disjoint from both S and O. We modify the tax policies $T_i^{\theta,\epsilon}$ only for agents i in Q, and otherwise, these policies are not altered. We assume that, for $i \in Q$, $T_i^{\theta,\epsilon}(z_i) = \bar{\tau}(\theta,\epsilon) z_i + \bar{\kappa}_i(\theta,\epsilon)$, where $\bar{\tau}(\theta,\epsilon)$ is a marginal tax rate, common to agents in Q, and $\bar{\kappa}_i(\theta,\epsilon)$ is a lump-sum tax. We may assume that, for each agent $i \in Q$, the lump-sum tax $\bar{\kappa}_i(\theta,\epsilon)$ is chosen so as to offset any utility change as the marginal tax rate $\bar{\tau}(\theta,\epsilon)$ changes, so that agents in Q are indifferent among tax policies $T^{\theta,\epsilon}$ as θ and ϵ vary. Note, however, that if the marginal tax rate changes, and the lump-sum tax adjusts to keep agents' utility constant, this will change the revenue raised by the tax policy. We may then also assume that $\bar{\tau}(\theta,\epsilon)$ (which determines $\kappa_i(\theta,\epsilon)$ for each i in Q up to a constant) is chosen so that the change in revenue among agents in Q just offsets any change in revenue among agents in S and O as θ and ϵ change. In this way, we keep revenue constant as we create the social preference cycle. The above arguments establishing the cycle are unaltered because agents in Q are indifferent as θ and ϵ change. A formal version of the Proof of Theorem 2 is in the online Appendix, and online Appendix A.8.2 shows how to fill in details when constructing $\left(T^{\theta,\epsilon}\right)$ so that it is well behaved.

B. *A Detailed Example: Libertarian Weights*

I now present a detailed example. The argument is parallel to that in the previous section, although some of the details differ. In particular, in this example, I no longer assume that taxes can be completely individualized. Instead, I assume that agents have a single observable binary characteristic x_i that takes values A and B. For $i \in [0,1/2]$, $x_i = A$, and for $i \in (1/2,1]$, $x_i = B$, so half of the population has

each characteristic. I assume it is possible to condition taxes on the characteristic, but the characteristic is not relevant to payoffs or welfare weights. In particular, all types share the same cost of earning income $v(z_i, A) = v(z_i, B) = v(z_i) = z_i^2/2$. I assume that welfare weights are libertarian and identical across agents, so that, for all $i \in [0,1]$, welfare weights are of the form $g_i(c_i, z_i) = \tilde{g}(t_i)$, where $\tilde{g}$ is increasing in the tax $t_i = z_i - c_i$ paid by the agent.

PROPOSITION 4: *In the model of the preceding paragraph, welfare weights are inconsistent in the sense there exist tax policies T_0, T_1, T_2, T_3, each of which raises the same revenue, and such that welfare weights imply a social preference cycle of the form $T_0 \prec^g T_1 \sim^g T_2 \prec^g T_3 \sim^g T_0$.*

This result resembles Theorem 2. Libertarian weights are not structurally utilitarian. This can be seen in the argument below, in which we construct a tax policy such that utility is held fixed but the total tax paid by specific agents, and hence also their libertarian welfare weight, varies.

I now establish the proposition. Consider linear taxes of the form $T(z) = \tau z + \kappa$, where τ is the marginal tax rate and κ is a lump-sum payment. Agents facing marginal tax rate τ solve the problem $\max_z \left[z(1-\tau) - z^2/2 - \kappa\right]$, and the optimal income is $z(\tau) = 1 - \tau$. Because utility is quasi-linear, $z(\tau)$ does not depend on the lump-sum tax. Define $\kappa(\tau)$ to be the lump-sum tax that makes agents' utility equal to zero when facing marginal tax rate τ (using the utility representation $\hat{U}_i(c_i, z_i) = c_i - v_i(z_i)$ that omits the outer utility function $u(\cdot)$). Formally, $\kappa(\tau)$ solves $z(\tau)(1-\tau) - \kappa(\tau) - v(z(\tau)) = 0$. Given our assumptions, $\kappa(\tau) = (1-\tau)^2/2$.[10] Consider the doubly parameterized family of tax policies $(T^{\theta,\epsilon})$, where $\theta \in [\theta_0, \theta_1]$, with $\theta_0 = \sqrt{1/3}, \theta_1 = \sqrt{2/3}$:

$$T_i^{\theta,\epsilon}(z_i) = \begin{cases} \theta z_i + \kappa(\theta) + \epsilon, & \text{if } x_i = A, \\ \left(\sqrt{1-\theta^2}\right) z_i + \kappa\left(\sqrt{1-\theta^2}\right) - \epsilon, & \text{if } x_i = B. \end{cases}$$

Observe first that ϵ just parameterizes a transfer from agents with characteristic A to agents with characteristic B; since utility is quasi-linear, such a transfer does not lead to a behavioral response, and hence, because there is an equal mass of type A and type B agents, the transfer is revenue neutral. Agents with characteristic A face a marginal tax rate of θ, and agents with characteristic B face a marginal tax rate of $\sqrt{1-\theta^2}$. As θ rises from θ_0 to θ_1, the marginal tax rate of type A agents rises from $\sqrt{1/3}$ to $\sqrt{2/3}$, while the marginal tax rate of type B agents falls from $\sqrt{2/3}$ to $\sqrt{1/3}$. Moreover, as θ rises from θ_0 to θ_1, the per agent revenue raised from type A agents falls from $1/3 + \epsilon$ to $1/6 + \epsilon$, and the per agent revenue raised from type B agents rises from $1/6 - \epsilon$ to $1/3 - \epsilon$.[11] The formula $\sqrt{1-\theta^2}$ was chosen for type B agents' marginal tax rate because this is the formula required for the revenue

[10] We have $\kappa(\tau) = z(\tau)(1-\tau) - v(z(\tau)) = (1-\tau)(1-\tau) - (1-\tau)^2/2 = (1-\tau)^2/2$.
[11] These numbers are derived in online Appendix A.9.

effects from type A and type B agents to exactly offset one another so that the total revenue of the tax policy remains equal to $1/4$ for all θ and ϵ.[12]

When $\epsilon = 0$, observe that the lump-sum tax $\kappa(\theta)$ is chosen so as to keep type A agents' utility equal to zero as θ varies. So type A agents are indifferent among all tax policies of the form $T^{\theta,0}$. Likewise, the lump-sum tax $\kappa\left(\sqrt{1-\theta^2}\right)$ makes type B agents indifferent among all tax policies of the form $T^{\theta,0}$. Because utility is quasi-linear, these indifference conditions continue to hold if, in addition, there is a fixed transfer from type A to type B agents. So for any fixed ϵ, all agents are indifferent among tax policies $T^{\theta,\epsilon}$ as θ varies. So by part (i) of Proposition 2—Pareto indifference along paths—it follows that varying θ from θ_0 to θ_1 is socially indifferent: $\forall \epsilon,\ T^{\theta_0,\epsilon} \sim^g T^{\theta_1,\epsilon}$.

As mentioned above, when $\theta = \theta_0$ and $\epsilon = 0$, type A agents pay a per person tax of $1/3$, while type B agents pay $1/6$. Since libertarian weights $\tilde{g}(t)$ are increasing in taxes paid t, half of the agents fall into each category A and B, and type A agents pay more in tax than type B agents, a small transfer $\epsilon > 0$ from type A to type B agents is *bad* at $T^{\theta_0,0}$ according to libertarian weights. That is, $T^{\theta_0,0} \succ^g T^{\theta_0,\epsilon}$ for sufficiently small $\epsilon > 0$. When $\theta = \theta_1$ and $\epsilon = 0$, the situation is exactly reversed, so that type A agents pay a tax of $1/6$, while type B agents pay $1/3$. So, at $T^{\theta_1,0}$, a small transfer from type A to type B is *good*. That is, $T^{\theta_1,0} \prec^g T^{\theta_1,\epsilon}$ for sufficiently small $\epsilon > 0$.[13]

Putting together the social preferences and indifferences derived in the preceding paragraphs, for sufficiently small $\epsilon > 0$, we have $T^{\theta_1,0} \prec^g T^{\theta_1,\epsilon} \sim^g T^{\theta_0,\epsilon} \prec^g T^{\theta_0,0} \sim^g T^{\theta_1,0}$. This establishes that the libertarian welfare weights imply a cycle. As I show in the next section, the fact that in this example, taxes depend on characteristics, specifically ones that do not affect utility, is inessential to the argument. The problem arises because endogenously chosen quantities (in this case $t = z - c$) can affect welfare weights without affecting utility.

V. The Main Theorem without Individualized Taxes

In this section, I show that it is possible to generate social preference cycles when all agents face the same tax schedule. The argument then becomes more complicated, but its overall structure is similar. One of the reasons that the proof becomes more complicated is that if taxes are not individualized, it will no longer be possible to hold all agents indifferent as we modify taxes in a nontrivial way. So the proof of the theorem in the general case no longer appeals to the Pareto indifference principle inherent in the welfare weights approach (Proposition 2). The step in the preceding argument in which all agents are kept indifferent as the parameter θ varies is replaced by a step in which, if benefits and costs to different agents are aggregated according to the system of social welfare weights g, then the change as θ varies is socially indifferent.

[12] This calculation is verified in online Appendix A.9.

[13] A formal derivation, appealing to Proposition 1, is in online Appendix A.9.

A. *Additional Assumptions*

For the main result, I assume that there are no observable characteristics, but there is a single one-dimensional real-valued unobservable characteristic y. Because there are no observable characteristics on which to condition taxes, I omit the subscript i on taxes and write $T(z_i)$ rather than $T_i(z_i)$. This also simplifies the definition of a well-behaved family of tax policies in Section IIA; condition (ii) in the definition of a well-behaved family $\left(T^{\theta}\right)$ simplifies to the map $(z,\theta) \mapsto T(z,\theta)$ is smooth (and similarly to $(z,\theta,\epsilon) \mapsto T(z,\theta,\epsilon)$ is smooth for a doubly parameterized family $\left(T^{\theta,\epsilon}\right)$; see online Appendix A.1.2 for a complete definition). I assume that the function $i \mapsto y_i$ assigning to each i their characteristic y_i is smooth, strictly increasing in i, and, more specifically, the derivative of y_i with respect to i is positive at all values of i in $I = [0,1]$. In this case we can write $v_i(z_i) = v(z_i, y_i)$ and $g_i(c_i, z_i) = g(c_i, z_i, y_i)$. Moreover, I assume that a higher value of y corresponds to the ability to earn income at a lower cost, so that $\forall z,\ \forall y, \frac{\partial^2}{\partial y \partial z} v(z,y) < 0$. This implies that, in response to any regular tax policy,[14] agents with a higher index i—hence a higher value of y_i—earn higher income.

B. *Statement of the Theorem*

THEOREM 3: *Under the supplementary assumptions of Section VA, if welfare weights g are not structurally utilitarian, then there exist tax policies* T_0, T_1, T_2, T_3, *each of which raises the same revenue, and such that welfare weights imply a social preference cycle of the form* $T_0 \prec^g T_1 \sim^g T_2 \prec^g T_3 \sim^g T_0$.

Together, Theorems 1 and 3 characterize the exact property on welfare weights—structural utilitarianism—that is required for welfare weights to be consistent. If welfare weights are structurally utilitarian, they are compatible with a social welfare function and hence with a consistent social preference, and if welfare weights are not structurally utilitarian, they imply a social preference cycle. This means that to acquire a *consistent* method of evaluating tax policies from welfare weights, generalized welfare weights must be quite similar to traditional welfare weights, and the promise of the GSMWW approach that one can represent very general values with generalized welfare weights is not fulfilled. To really represent broader values, we need to seek more general approaches that differ more fundamentally from the traditional utilitarian approach.

C. *Proof Sketch*

Here, I sketch the proof of the main theorem; the missing details can be found in the online Appendix. Like in the proof of the simpler version of the theorem in Section IVA, we construct a doubly parameterized family of tax policies, $\left(T^{\theta,\epsilon}\right)_{\theta\in\Theta,\epsilon\in E}$, where $\Theta = \left[\underline{\theta}, \overline{\theta}\right]$ and $E = \left[\underline{\epsilon}, \overline{\epsilon}\right]$. Heuristically, we can think

[14] Assuming that taxes are not individualized also simplifies the characterization of regular tax policies; see online Appendix A.1.2.

of ϵ as parameterizing a redistribution from some set of higher-income agents S to a set of lower-income agents O; as ϵ rises, taxes on agents in S rise, while those in O fall. The specific construction of $T^{\theta,\epsilon}$ in the online Appendix bears out this interpretation (see the Proof of Lemma 3), and, in this way, the argument resembles the argument in Section IVA.

Sufficient Conditions for a Social Preference Cycle.—Now suppose that we construct such a family $\left(T^{\theta,\epsilon}\right)$ with the following two properties:

Indifference to θ: Holding fixed ϵ, the value of θ is socially indifferent:

$$\forall \epsilon \in E, \quad \forall \theta' \in \Theta, \quad \int g_i(\theta', \epsilon) \frac{\partial}{\partial \theta}\Big|_{\theta=\theta'} T\big(z_i(\theta', \epsilon), \theta, \epsilon\big) di = 0. \tag{15}$$

Changing Desirability of Redistribution ϵ: There exist $\theta_0 \in \left(\underline{\theta}, \bar{\theta}\right)$ and $\epsilon_0 \in \left(\underline{\epsilon}, \bar{\epsilon}\right)$ such that at (θ_0, ϵ_0), as θ crosses θ_0, a change in ϵ goes from being undesirable to being desirable:

$$\int g_i(\theta_0, \epsilon_0) \frac{\partial}{\partial \epsilon}\Big|_{\epsilon=\epsilon_0} T\big(z_i(\theta_0, \epsilon_0), \theta_0, \epsilon\big) di = 0, \tag{16}$$

$$\frac{d}{d\theta}\Big|_{\theta=\theta_0} \int g_i(\theta, \epsilon_0) \frac{\partial}{\partial \epsilon}\Big|_{\epsilon=\epsilon_0} T\big(z_i(\theta, \epsilon_0), \theta, \epsilon\big) di < 0. \tag{17}$$

The following lemma shows that, if we can construct a family $\left(T^{\theta,\epsilon}\right)$ with the above properties, that is sufficient to construct a social preference cycle.

LEMMA 1: *Suppose that welfare weights g are such that there exists a well-behaved family* $\left(T^{\theta,\epsilon}\right)$ *that satisfies* (15)–(17). *Then there exist parameter values* $\theta_-, \theta_+ \in \Theta$ *and* $\epsilon_0, \epsilon_+ \in E$ *for which there exists a social preference cycle of the form* $T^{\theta_+,\epsilon_0} \prec^g T^{\theta_+,\epsilon_+} \sim^g T^{\theta_-,\epsilon_+} \prec^g T^{\theta_-,\epsilon_0} \sim^g T^{\theta_+,\epsilon_0}$.

PROOF:

Suppose there is a family $\left(T^{\theta,\epsilon}\right)$ satisfying (15)–(17). Then (16) and (17) imply that for $\theta_- \in \Theta$ such that $\theta_- < \theta_0$ and θ_- is sufficiently close to θ_0, $\int g_i(\theta_-, \epsilon_0) \frac{\partial}{\partial \epsilon}\Big|_{\epsilon=\epsilon_0} T\big(z_i(\theta_-, \epsilon_0), \theta_-, \epsilon\big) di > 0$, while at the same time for $\theta_+ \in \Theta$ such that $\theta_+ > \theta_0$ and θ_+ is sufficiently close to θ_0, $\int g_i(\theta_+, \epsilon_0) \frac{\partial}{\partial \epsilon}\Big|_{\epsilon=\epsilon_0} T\big(z_i(\theta_+, \epsilon_0), \theta_+, \epsilon\big) di < 0$. It follows from these two inequalities and the local improvement principle (Proposition 1) that for $\epsilon_+ \in E$ such that $\epsilon_+ > \epsilon_0$ and ϵ_+ sufficiently close to ϵ_0, $T^{\theta_-,\epsilon_0} \succ^g T^{\theta_-,\epsilon_+}$ and $T^{\theta_+,\epsilon_0} \prec^g T^{\theta_+,\epsilon_+}$. It follows from (15) and the global indifference principle (Section IIB) that $T^{\theta_-,\epsilon_0} \sim^g T^{\theta_+,\epsilon_0}$ and $T^{\theta_-,\epsilon_+} \sim^g T^{\theta_+,\epsilon_+}$.[15] Putting the just derived relations together, we derive the cycle promised by the lemma. ∎

[15] Observe that when $\left(T^{\theta,\epsilon}\right)$ is well behaved, then, for each fixed $\epsilon \in E$, the family $\left(T^{\theta,\epsilon}\right)_{\theta\in\Theta}$ is well behaved, and for each $\theta \in \Theta$, $\left(T^{\theta,\epsilon}\right)_{\epsilon\in E}$ is well behaved. So the improvement and indifference principles can be applied to one of the parameters θ or ϵ at a time, holding the other fixed.

Nonstructurally Utilitarian Weights Allow a Family $\left(T^{\theta,\epsilon}\right)$ *Satisfying the Sufficient Conditions for a Social Preference Cycle*.—I now show that the sufficient conditions for a social preference cycle (15)–(17) are jointly satisfiable if (and only if) welfare weights are not structurally utilitarian. It is convenient to define $\hat{U}_i(T) = \hat{U}_i\big(c_i(T), z_i(T)\big)$ and $\hat{U}_i(\theta,\epsilon) = \hat{U}_i\big(T^{\theta,\epsilon}\big)$. I begin by stating a fairly immediate corollary of Proposition 3, which is proved in the online Appendix.

COROLLARY 3: *If g is not structurally utilitarian, then there exists a regular tax policy T for which there exist agents* $i_a, i_b \in (0,1)$ *with* $i_a < i_b$ *such that either*

$$\forall\, i \in (i_a, i_b),\ \frac{\partial}{\partial z_i}\hat{g}_i\big(\hat{U}_i(T), z_i(T)\big) < 0 \tag{18}$$

or

$$\forall\, i \in (i_a, i_b),\ \frac{\partial}{\partial z_i}\hat{g}_i\big(\hat{U}_i(T), z_i(T)\big) > 0. \tag{19}$$

Next, I show that in the presence of condition (15), condition (17) takes a more convenient form.

LEMMA 2: *Assume that* $\left(T^{\theta,\epsilon}\right)$ *is well behaved and satisfies* (15). *Then* (17) *holds if and only if*

$$\begin{aligned}\int \frac{\partial}{\partial z_i}\hat{g}_i\big(\hat{U}_i(\theta_0,\epsilon_0), z_i(\theta_0,\epsilon_0)\big)\Bigg[&\frac{\partial}{\partial\theta}\bigg|_{\theta=\theta_0} z_i(\theta,\epsilon_0)\frac{\partial}{\partial\epsilon}\bigg|_{\epsilon=\epsilon_0} T\big(z_i(\theta_0,\epsilon_0),\theta_0,\epsilon\big)\\ &- \frac{\partial}{\partial\epsilon}\bigg|_{\epsilon=\epsilon_0} z_i(\theta_0,\epsilon)\frac{\partial}{\partial\theta}\bigg|_{\theta=\theta_0} T\big(z_i(\theta_0,\epsilon_0),\theta,\epsilon_0\big)\Bigg] di\\ &< 0.\end{aligned} \tag{20}$$

Since, by Proposition 3, for structurally utilitarian weights, $\frac{\partial}{\partial z_i}\hat{g}_i\big(\hat{U}_i(\theta_0,\epsilon_0), z_i(\theta_0,\epsilon_0)\big) = 0$ everywhere, the integral in the left-hand side of (20) is always equal to zero when welfare weights are structurally utilitarian. Hence, it follows immediately from Lemma 2 that a necessary condition for (15)–(17) to be satisfied is for welfare weights *not* to be structurally utilitarian. However, what we need to show here is that not being structurally utilitarian is a *sufficient* condition for the ability to construct a family of tax policies for which (15)–(17) to hold.

PROOF OUTLINE OF LEMMA 2:

The key is to show that, when expanded, the expression in the left-hand side of (17) and the ϵ-derivative of the expression on the left-hand side of (15), evaluated

at (θ_0, ϵ_0), have overlapping terms. In particular, I will define terms A, B, and C such that

$$\frac{d}{d\epsilon}\Big|_{\epsilon=\epsilon_0} \int g_i(\theta_0, \epsilon) \frac{\partial}{\partial \theta}\Big|_{\theta=\theta_0} T\big(z_i(\theta_0, \epsilon), \theta, \epsilon\big) di \;=\; A + C, \tag{21}$$

$$\frac{d}{d\theta}\Big|_{\theta=\theta_0} \int g_i(\theta, \epsilon_0) \frac{\partial}{\partial \epsilon}\Big|_{\epsilon=\epsilon_0} T\big(z_i(\theta, \epsilon_0), \theta, \epsilon\big) di \;=\; B + C. \tag{22}$$

Above,

$$A \;=\; \int \frac{\partial}{\partial z_i} \hat{g}_i\big(\hat{U}_i(\theta_0, \epsilon_0), z_i(\theta_0, \epsilon_0)\big) \frac{\partial}{\partial \epsilon}\Big|_{\epsilon=\epsilon_0} z_i(\theta_0, \epsilon) \frac{\partial}{\partial \theta}\Big|_{\theta=\theta_0} T\big(z_i(\theta_0, \epsilon_0), \theta, \epsilon_0\big) di,$$

$$B \;=\; \int \frac{\partial}{\partial z_i} \hat{g}_i\big(\hat{U}_i(\theta_0, \epsilon_0), z_i(\theta_0, \epsilon_0)\big) \frac{\partial}{\partial \theta}\Big|_{\theta=\theta_0} z_i(\theta, \epsilon_0) \frac{\partial}{\partial \epsilon}\Big|_{\epsilon=\epsilon_0} T\big(z_i(\theta_0, \epsilon_0), \theta_0, \epsilon\big) di.$$

The term C, as well as the derivation of (21) and (22), are in the online Appendix. Note that (15) implies that the left-hand side of (21) is equal to zero, which implies that the right-hand side is equal to zero as well. It follows that $C = -A$. So $B + C = B - A$. It follows that the left-hand side of (22) is less than zero, which is what (17) says, if and only if $B - A < 0$. But $B - A < 0$ is equivalent to (20). This completes the Proof of Lemma 2. ∎

The next lemma shows that in order to be able to construct a family $\big(T^{\theta,\epsilon}\big)$ that satisfies (15), (16), and (20), it is sufficient to find a tax policy T and i_a, i_b for which (19) holds. (The online Appendix presents an analogous lemma, Lemma A.2, corresponding to condition (18).)

LEMMA 3: *Let T be a regular tax policy, and let $i_a, i_b \in (0,1)$ be such that $i_a < i_b$. Then there exists a well-behaved family $\big(T^{\theta,\epsilon}\big)$ with $T^{\theta_0,\epsilon_0} = T$ for some interior parameter values θ_0, ϵ_0 and that satisfies* (15), (16), *and*

$$\frac{\partial}{\partial \theta}\Big|_{\theta=\theta_0} z_i(\theta, \epsilon_0) \frac{\partial}{\partial \epsilon}\Big|_{\epsilon=\epsilon_0} T\big(z_i(\theta_0, \epsilon_0), \theta_0, \epsilon\big) \tag{23}$$

$$-\frac{\partial}{\partial \epsilon}\Big|_{\epsilon=\epsilon_0} z_i(\theta_0, \epsilon) \frac{\partial}{\partial \theta}\Big|_{\theta=\theta_0} T\big(z_i(\theta_0, \epsilon_0), \theta, \epsilon_0\big) \begin{cases} <0, & \text{if } i \in (i_a, i_b) \\ =0, & \text{if } i \notin (i_a, i_b) \end{cases}.$$

The lemma is proven in the online Appendix. This lemma does not depend on any assumptions on welfare weights but just on the broad flexibility that is available in constructing tax policies.

The following lemma puts together the previous results derived in this section.

LEMMA 4: *If g is not structurally utilitarian, then there exists a well-behaved family of tax policies* $\left(T^{\theta,\epsilon}\right)$ *satisfying* (15)–(17).

PROOF:

Assume that g is not structurally utilitarian. It then follows from Corollary 3 that there exists a regular tax policy T and $i_a, i_b \in (0,1)$ such that $i_a < i_b$ and either (18) or (19) hold. First assume that (19) holds. It follows from Lemma 3 that there exists a well-behaved family of tax policies $\left(T^{\theta,\epsilon}\right)_{\theta\in\Theta,\epsilon\in E}$, with $T^{\theta_0,\epsilon_0} = T$ satisfying (15), (16), and (23), where, in (23), i_a and i_b are chosen to be the same values for which (19) holds. Moreover, (19) and (23) together imply (20). So in this case, we can construct well-behaved family $\left(T^{\theta,\epsilon}\right)$ satisfying (15), (16), and (20). A similar argument—invoking a variant of Lemma 3 (Lemma A.2 in online Appendix A.12.3)—shows that, when (18) rather than (19) holds, we can still construct a well-behaved family $\left(T^{\theta,\epsilon}\right)$ satisfying (15), (16), and (20). It now follows from Lemma 2 that whenever welfare weights are not structurally utilitarian, it is possible to construct a tax policy satisfying (15)–(17). ∎

Holding Revenue Constant.—The construction of the previous section can be extended so that the family $\left(T^{\theta,\epsilon}\right)$ is chosen to hold revenue constant, as stated by the following lemma.

LEMMA 5: *If g is not structurally utilitarian, then there exists a well-behaved constant revenue family of tax policies* $\left(T^{\theta,\epsilon}\right)$ *satisfying* (15)–(17).

Lemma 5 is a strengthening of Lemma 4 that differs from Lemma 4 only in that family $\left(T^{\theta,\epsilon}\right)$ is required to be a constant-revenue family in the sense that all tax policies $T^{\theta,\epsilon}$ raise the same revenue. I have separated this additional requirement into a separate lemma because the argument that revenue can be held constant appeals to different principles than the proof of the other properties. The basic idea is similar to that described in Section IVA for holding revenue constant. In particular, once we construct a family $\left(T^{\theta,\epsilon}\right)$ satisfying (15)–(17), as we know we can do from Lemma 4, we consider a positive measure set Q of agents at a different income level than agents in S and O and vary the revenue raised from agents in Q as θ and ϵ vary exactly so as to offset revenue changes elsewhere in the tax schedule in such a way that there is no detectable welfare change in Q according to welfare weights; this is analogous to moving along a social indifference curve for agents in Q along which the revenue raised from those agents varies. The details are in the online Appendix.

Putting It All Together.—Putting Lemmas 1 and 5 together yields Theorem 3, the main result.

D. *An Application: Poverty Alleviation*

I now present an application to illustrate the main result. Maintain all of the assumptions of Section VA. Let $\bar{c}$ be the poverty line; that is, $\bar{c}$ is the level of consumption

below which agents are considered to be poor. Now consider welfare weights that capture the goal of poverty alleviation by concentrating weight on agents beneath the poverty line. Saez and Stantcheva (2016) presented such an example.[16] I modify their example slightly to make welfare weights smooth. Suppose that $g_i(c_i, z_i) = \tilde{g}(c_i)$, where $\tilde{g}(c_i)$ is decreasing in c_i until c_i gets to $\bar{c}$ and then remains constant at the value $\underline{g}$ thereafter, where $\underline{g} > 0$. I assume that $\underline{g} > 0$ to be in conformity with my prior assumptions, but we may assume that $\underline{g}$ is arbitrarily close to zero. So agents below the poverty line have a higher welfare weight than agents above the poverty line, the welfare weight is greater the further below the poverty line the agent is, and it is constant for agents above the poverty line.

Now consider a doubly parameterized family of tax policies $\left(T^{\theta,\epsilon}\right)$ of the form $T(z,\theta,\epsilon) = \theta f(z) + (\theta - \epsilon)z + \alpha\epsilon - \kappa(\theta,\epsilon)$, where $f(z)$ is a smooth function and, for some θ_0, $\kappa(\theta_0,\epsilon) = 0, \forall\epsilon$. Assume that there exists ϵ_0 and income level $\bar{z}$ (within the income distribution), such that, when facing tax schedule T^{θ_0,ϵ_0}, all agents earn positive income, all agents earning income $\bar{z}$ or above are strictly above the poverty line, and a positive measure of agents with income below $\bar{z}$ are beneath the poverty line. I assume that $f(z) = 0$ for all z with $z \leq \bar{z}$, and $f(z) > 0$ for all z with $z > \bar{z}$, so that the $\theta f(z)$ term specifies taxes that only apply to agents above the poverty line when (θ,ϵ) is close to (θ_0,ϵ_0). Noting that the optimal income for i, $z_i(\theta,\epsilon)$, is independent of α, assume that α is chosen so that $\int g(\theta_0,\epsilon_0)\left[z_i(\theta_0,\epsilon_0) - \alpha\right]di = 0$, which says that, at T^{θ_0,ϵ_0}, the positive welfare effect of increasing ϵ due to decreasing marginal tax rates through the term $-\epsilon z$ is just offset by the negative welfare effect of the increase in the lump-sum tax $\alpha\epsilon$. (Note that, by our assumptions, $\frac{\partial}{\partial\epsilon}\Big|_{\epsilon=\epsilon_0}\kappa(\theta_0,\epsilon) = 0$.) Finally, we assume that $\kappa(\theta,\epsilon)$ satisfies the following set of differential equations (note that $g_i(\theta,\epsilon)$ depends on $\kappa(\theta,\epsilon)$):

$$\frac{\partial}{\partial\theta}\Big|_{\theta=\theta'}\kappa(\theta,\epsilon) = \int \frac{g_i(\theta',\epsilon)}{\int g_j(\theta',\epsilon)dj}\left[z_i(\theta',\epsilon) + f\left(z_i(\theta',\epsilon)\right)\right]di, \ \forall\theta', \ \forall\epsilon. \tag{24}$$

Rearranging terms, one can see that (24) says that for any fixed value of ϵ, when changing θ, the welfare effect due to increasing marginal tax rates through the term $\theta f(z) + \theta z$ is just offset by the welfare effect of the change in the lump-sum tax $\kappa(\theta,\epsilon)$. Note that the differential equations (24) and the conditions $\kappa(\theta_0,\epsilon) = 0$, $\forall\epsilon$ uniquely determine $\kappa(\theta,\epsilon)$.

PROPOSITION 5: *With poverty alleviation welfare weights, if* $\left(T^{\theta,\epsilon}\right)$ *has the properties assumed in this section,* $\left(T^{\theta,\epsilon}\right)$ *satisfies* (15)–(17), *the sufficient conditions for a preference cycle in Lemma 1.*

[16] Besley and Coate (1992) and Kanbur, Keen, and Tuomala (1994) incorporate poverty alleviation in optimal tax.

The proof is in the online Appendix. Condition (16) corresponds to $\int g(\theta_0,\epsilon_0)\big[z_i(\theta_0,\epsilon_0)-\alpha\big]di = 0$ (and $\frac{\partial}{\partial\epsilon}\big|_{\epsilon=\epsilon_0}\kappa(\theta_0,\epsilon) = 0$), and (15) corresponds to (24). The key calculation that drives the argument is that

$$\frac{d}{d\theta}\Big|_{\theta=\theta_0}\int g_i(\theta,\epsilon_0)\frac{\partial}{\partial\epsilon}\Big|_{\epsilon=\epsilon_0} T\big(z_i(\theta,\epsilon_0),\theta,\epsilon\big)di$$

$$= \int \tilde{g}'\big(c_i(\theta_0,\epsilon_0)\big)\frac{v_i'\big(z_i(\theta_0,\epsilon_0)\big)}{v_i''\big(z_i(\theta_0,\epsilon_0)\big)}di \times \int \frac{\tilde{g}\big(c_i(\theta_0,\epsilon_0)\big)}{\int \tilde{g}\big(c_j(\theta_0,\epsilon_0)\big)dj} f\big(z_i(\theta_0,\epsilon_0)\big)di < 0,$$

which establishes (17). It follows from Proposition 5 that, in the poverty alleviation example, with tax policies as described above, we can construct a social preference cycle exactly as in the Proof of Lemma 1 (see Section VC). We have not worried about holding revenue constant, but Lemma 5 tells us that we can modify the construction of $\big(T^{\theta,\epsilon}\big)$ so as to hold revenue constant as well. Of course, the reason we could construct a cycle is that poverty reduction welfare weights are not structurally utilitarian. In particular, by increasing both consumption c_i and income z_i so as to hold total utility $u\big(c_i - v_i(z_i)\big)$ fixed, it is possible to bring an agent above the poverty line, and, in this way, we can change their welfare weight; this is not consistent with structural utilitarianism. In general, welfare weights that respond to changes in consumption but do not take into account labor supply costs will not be structurally utilitarian and hence will lead to social preference cycles. More generally, welfare weights that respond to only a subset of the endogenously chosen arguments that determine utility will be vulnerable to inconsistency.

VI. Generalization to Non-quasi-linear Preferences

Throughout the paper, I assumed quasi-linear utility, which rules out income effects. This section discusses how the results generalize without quasi-linearity. For more general utility functions $U_i(c_i,z_i)$ that are not necessarily quasi-linear, structural utilitarianism can be defined as follows.

DEFINITION 2 (Structural Utilitarianism without Quasi-Linearity)*: A system of welfare weights g is* ***structurally utilitarian*** *if and only if* $\forall\, i \in I,\ \forall\, z_i, z_i' \in Z,\ \forall\ c_i, c_i' \in \mathbb{R}$,

$$U_i(c_i,z_i) = U_i(c_i',z_i') \Rightarrow \frac{\frac{\partial}{\partial c_i}U_i(c_i,z_i)}{\frac{\partial}{\partial c_i}U_i(c_i',z_i')} = \frac{g_i(c_i,z_i)}{g_i(c_i',z_i')}. \tag{25}$$

This condition says that, as we move along a fixed (c_i,z_i)-indifference curve for agent i, i's marginal welfare weight must be proportional to the marginal utility of consumption. Of course, utilitarian weights must satisfy this condition, as they are equal to the marginal utility of consumption. Section III provided several equivalent

conditions characterizing structural utilitarianism for the quasi-linear case, (6), (7), and (8). To see that (25) is indeed a generalization of these conditions, it is easiest to compare with (8). As discussed in Section III, for quasi-linear utility, $U_i(c_i, z_i) = U_i(c_i', z_i')$ implies $\frac{\partial}{\partial c_i} U_i(c_i, z_i) = \frac{\partial}{\partial c_i} U_i(c_i', z_i')$, or equivalently, if $U_i(c_i, z_i) = U_i(c_i', z_i')$, then $\frac{\partial}{\partial c_i} U_i(c_i, z_i) / \frac{\partial}{\partial c_i} U_i(c_i', z_i') = 1$. So, with quasi-linearity, (25) reduces to $U_i(c_i, z_i) = U_i(c_i', z_i') \Rightarrow 1 = g_i(c_i, z_i) / g_i(c_i', z_i')$, which is equivalent to (8). In other words, for quasi-linear utility, the marginal utility of consumption is constant along any (c_i, z_i)-indifference curve, and hence, (25) says that structurally utilitarian welfare weights must be constant too. So Definition 2 indeed generalizes the previous definition of structural utilitarianism.

Our results also generalize. Even without quasi-linearity, welfare weights are structurally utilitarian if and only if they arise from a generalized utilitarian social welfare function—so that Theorem 1 still holds—and if welfare weights are not structurally utilitarian, then it is possible to construct a social preference cycle so that Theorem 3 holds as well. These results assume some regularity conditions on the utility functions $U_i(c_i, z_i)$. Online Appendix D presents these conditions and explains how to modify the proofs of the theorems when quasi-linearity is no longer assumed.

VII. Discussion

The motivation for generalized social marginal welfare weights was as a means of addressing the omission of broader values in economic analysis. I have argued in this paper that this solution does not work because generalized welfare weights, once they stray too far from traditional utilitarian weights, are inconsistent. In this closing section, I will discuss some related literature and how the current contribution differs, as well as ways forward on the problem of incorporating broader normative values in economic analysis.

A. *The Pareto Principle and Broader Values: Related Literature*

Saez and Stantcheva (2016, p. 25) write, "if the weights are nonnegative, then our theory respects the Pareto principle in the sense that, around the local optimum, there is no Pareto improving small reform." It may appear that Saez and Stantcheva (2016) have uncovered a way of incorporating broader values into economic analysis compatibly with the Pareto principle. Several authors, including Sen (1970, 1979a,b) and Kaplow and Shavell (2001, 2002), have argued that incorporating broader moral considerations into economic evaluation is inconsistent with the Pareto principle. Sen (1970, 1979a,b) interprets this as an argument against insisting on the Pareto principle, whereas Kaplow and Shavell (2001) interpret it as an argument against including nonwelfarist considerations in normative economic evaluation.[17] As they say, one philosopher's modus ponens is another philosopher's modus tollens. Fleurbaey, Tungodden, and Chang (2003) are critical of Kaplow and Shavell

[17] See also Weymark (2017).

(2001) and take a more positive view of incorporating broader values compatibly with the Pareto principle. In discussing the Saez and Stantcheva (2016) approach critically, Fleurbaey and Maniquet (2018), who also take a more positive view of incorporating broader values, compatibly with Pareto, write,

> … the social welfare function approach has been introduced by Bergson (1938) and Samuelson (1947) not out of a taste for elegance, but because it is the only way to define social preferences that are both transitive and Paretian. Therefore, a method that directly weights tax changes at the various earning levels is compatible with transitive and Paretian social preferences, and then extendable to the study of nonlocal reforms, only if it relies on the classical framework of the social welfare function. (p. 1059)

This informal passage is closely related to the results developed formally in the current paper.

B. *The Contribution of This Paper*

My result differs from the Kaplow and Shavell (2001) result in three ways: (i) Kaplow and Shavell (2001) are concerned with *social welfare functions*, which give global rankings, while I am concerned with *systems of generalized welfare weights*, which give local marginal rates of substitution.[18] (ii) The key property for Kaplow and Shavell (2001) is whether a social welfare function is *individualistic*, meaning that changes in states that do not affect individual utility cannot affect social welfare, whereas the key property for me is *structural utilitarianism*. Structural utilitarianism, at least under the assumption of quasi-linear preferences, is thematically similar to individualism in that both say that some aspect of social evaluation cannot change in response to certain types of changes that do not affect individual utility, but formally, the two properties are quite different, imposing different restrictions on different types of formal objects.[19] This difference is perhaps easiest to see by observing that individualism is equivalent to the property of Pareto indifference and, by Proposition 2, the social preferences induced by welfare weights satisfy a version of Pareto indifference regardless of whether they are structurally utilitarian. (iii) For Kaplow and Shavell (2001), the penalty for violating their key property is that the social ranking *violates weak Pareto*, whereas, for me, the penalty is that the implied social ranking contains a preference cycle and hence is *inconsistent*. My paper shows that eschewing social welfare functions in favor of the local comparisons of generalized marginal welfare weights is not a successful approach to avoiding Kaplow-Shavell-type impossibilities because it leads to inconsistencies.

[18] I do, however, bridge the gap between social welfare functions and marginal welfare weights to some extent by showing how to derive some of the global comparisons implied by welfare weights.

[19] Individualism says that changes in *social states* that do not affect utility do not affect *social welfare*, whereas, under the assumption of quasi-linearity, structural utilitarianism says that changes in agent's *decisions* (specifically of consumption and income) that do not affect an individual's utility do not affect *that individual's welfare weight*, but structural utilitarianism allows that exogenous characteristics contained in (x_i, y_i) may affect welfare weights without affecting utility. Moreover, in the non-quasi-linear case, structural utilitarianism generalizes to the property that, along any (c_i, z_i)-indifference curve, an agent's welfare weight is proportional to their marginal utility of consumption, which does not seem to be analogous to individualism in the same way as in the quasi-linear case.

Fleurbaey and Maniquet (2018) only discuss the potential intransitivity of welfare weights briefly, and they do not present a formal result characterizing when generalized social welfare weights are consistent. Nor do they provide a methodology for collecting the local judgments of the generalized social welfare weights into implicit global comparisons. In this paper, I do both of these things. I show how to collect the local judgments of generalized social welfare weights into global social judgments (see Section IIB) and that the precise property that is necessary and sufficient for welfare weights to be consistent is structural utilitarianism (see Theorems 1 and 3). Unlike Fleurbaey and Maniquet (2018), I also construct specific examples of cases in which generalized welfare weights are inconsistent. Moreover, my result is stronger than the point made by Fleurbaey and Maniquet (2018) in another way. I show that when welfare weights are not structurally utilitarian, they are not consistent with *any* social welfare function, *Paretian or not*. Notice, in this regard, that Theorem 3 does not mention any Pareto principle; it simply says that if welfare weights are not structurally utilitarian, then they are inconsistent.

C. *Two Ways Forward*

I now highlight two ways forward if broader values are to be incorporated into normative economic analysis and specifically optimal tax. (Fleurbaey and Maniquet 2018, p. 1031) write that "the classical social welfare function framework is more flexible than commonly thought, and can accommodate a very large set of nonutilitarian values. More specifically, fairness concepts can help solve the interpersonal comparison difficulties that the utilitarian approach faces when agents have different preferences by providing useful selections of suitable individual utility indexes," and their paper shows that *Paretian* social welfare functions can capture a broad set of values in an optimal tax context.[20] In the setting of the current paper, Theorem 1 shows that structurally utilitarian welfare weights are compatible with a generalized utilitarian social welfare function of the form $\int F\big(U_i(c_i, z_i), x_i, y_i\big)di$. We may think of the function $F\big(u_i, x_i, y_i\big)$ as reweighting utilities u_i—and hence also reweighting the social value we assign to tax changes—on the basis of certain moral considerations that are responsive to the characteristics $\big(x_i, y_i\big)$. The welfare weights induced by such a social welfare function must be consistent because they are derived from a consistent social ranking to begin with.

Not all values can be captured with Paretian approaches.[21] The second way forward embraces this point. Consider libertarianism as an example.[22] Suppose that one thinks that people are entitled to their pretax incomes and that in some way taxation is like theft. This view is not faithfully rendered as saying that additional income to people who have been taxed more should be given additional weight in comparison

[20] Other work representing broader values with Paretian social welfare functions includes Fleurbaey and Maniquet (2011); Piacquadio (2017); and Berg and Piacquadio (2023).

[21] Fleurbaey and Maniquet (2018, p. 1040) recognize this, writing, "we highlight another way in which at least some fairness principles can remain compatible with the Pareto principle … Not all fairness principles fall in this category, obviously, and the socialist and libertarian principles mentioned two paragraphs earlier provide examples of non-Paretian approaches." For criticisms of the Pareto principle, see Sen (1979b); Mongin (2016); and Sher (2020).

[22] For approaches to libertarian taxation, see Nozick (1974); Feldstein (1976); Young (1987); Weinzierl (2014); and Vallentyne (2018). For an approach to nonwelfarist optimal taxation, see Kanbur, Pirttilä, and Tuomala (2006).

to those who have been taxed less; rather, it is the view that it is wrong to tax, or at least, if not absolutely wrong, that it is bad to tax, and that this bad is tolerated, to the extent that it is, because of the other important purposes of taxation. On a rights-based version of libertarianism, taxing people is bad not because it reduces their utility but because it violates their entitlements. Imagine there is a function $s(t_i)$ for each agent i that measures how bad it is to violate i's entitlements. We might then maximize the non-Paretian social welfare function $W(T) = -\int_i s\left(T_i\left(z_i(T)\right)\right)di$ subject to a revenue requirement. Such an approach will not be Paretian, even in the sense of Proposition 2, and so it follows from Corollary 1 that this approach cannot be captured by welfare weights. Alternatively, we may trade off rights-based concerns as captured by $s(t_i)$ against utilitarian concerns. Or we may want to go farther and consider more thoroughly procedural approaches that do not appeal to a social objective (or even a local social objective). Whatever the right approach, it seems unlikely that we can capture the richness of broader ethical values by means of conservative modifications, such as by modifications of welfare weights, in a way that strongly preserves the structure of traditional optimal tax theory; we should expect that incorporation of broader values will require a more thorough change in the way that we normatively evaluate taxes and other economic policies.

REFERENCES

Berg, Kristoffer, and Paolo G. Piacquadio. 2023. "Fairness and Paretian Social Welfare Functions." Unpublished.

Bergson, Abram. 1938. "A Reformulation of Certain Aspects of Welfare Economics." *Quarterly Journal of Economics* 52 (2): 310–34.

Besley, Timothy, and Stephen Coate. 1992. "Workfare versus Welfare: Incentive Arguments for Work Requirements in Poverty-Alleviation Programs." *American Economic Review* 82 (1): 249–61.

Feldstein, Martin. 1976. "On the Theory of Tax Reform." *Journal of Public Economics* 6 (1-2): 77–104.

Fleurbaey, Marc, and François Maniquet. 2011. *A Theory of Fairness and Social Welfare.* Cambridge, UK: Cambridge University Press.

Fleurbaey, Marc, and François Maniquet. 2018. "Optimal Income Taxation Theory and Principles of Fairness." *Journal of Economic Literature* 56 (3): 1029–79.

Fleurbaey, Marc, Bertil Tungodden, and Howard F. Chang. 2003. "Any Non-welfarist Method of Policy Assessment Violates the Pareto Principle: A Comment." *Journal of Political Economy* 111 (6): 1382–85.

Kanbur, Ravi, Jukka Pirttilä, and Matti Tuomala. 2006. "Non-welfarist Optimal Taxation and Behavioural Public Economics." *Journal of Economic Surveys* 20 (5): 849–68.

Kanbur, Ravi, Michael Keen, and Matti Tuomala. 1994. "Optimal Non-linear Income Taxation for the Alleviation of Income-Poverty." *European Economic Review* 38 (8): 1613–32.

Kaplow, Louis, and Steven Shavell. 2001. "Any Non-welfarist Method of Policy Assessment Violates the Pareto Principle." *Journal of Political Economy* 109 (2): 281–86.

Kaplow, Louis, and Steven Shavell. 2002. *Fairness versus Welfare.* Cambridge, MA: Harvard University Press.

Mirrlees, James A. 1971. "An Exploration in the Theory of Optimum Income Taxation." *Review of Economic Studies* 38 (2): 175–208.

Mongin, Philippe. 2016. "Spurious Unanimity and the Pareto Principle." *Economics and Philosophy* 32 (3): 511–32.

Nozick, Robert. 1974. *Anarchy, State, and Utopia.* New York, NY: Basic Books.

Piacquadio, Paolo Giovanni. 2017. "A Fairness Justification of Utilitarianism." *Econometrica* 85 (4): 1261–76.

Saez, Emmanuel, and Stefanie Stantcheva. 2016. "Generalized Social Marginal Welfare Weights for Optimal Tax Theory." *American Economic Review* 106 (1): 24–45.

Samuelson, Paul Anthony. 1947. *Foundations of Economic Analysis.* Cambridge, MA: Harvard University Press.

Sen, Amartya. 1970. "The Impossibility of a Paretian Liberal." *Journal of Political Economy* 78 (1): 152–57.

Sen, Amartya. 1979a. "Personal Utilities and Public Judgements: or What's Wrong with Welfare Economics." *Economic Journal* 89 (355): 537–58.

Sen, Amartya. 1979b. "Utilitarianism and Welfarism." *Journal of Philosophy* 76 (9): 463–89.

Sher, Itai. 2020. "How Perspective-based Aggregation Undermines the Pareto Principle." *Politics, Philosophy and Economics* 19 (2): 182–205.

Vallentyne, Peter. 2018. "Libertarianism and Taxation." In *Taxation: Philosophical Perspectives*, edited by Martin O'Neill and Shepley Orr, 98–110. Oxford: Oxford University Press.

Weinzierl, Matthew. 2014. "The Promise of Positive Optimal Taxation: Normative Diversity and a Role for Equal Sacrifice." *Journal of Public Economics* 118: 128–42.

Weinzierl, Matthew. 2017. "Popular Acceptance of Inequality due to Innate Brute Luck and Support for Classical Benefit-based Taxation." *Journal of Public Economics* 155: 54–63.

Weymark, John A. 2017. "Conundrums for Nonconsequentialists." *Social Choice and Welfare* 48 (2): 269–94.

Young, H. Peyton. 1987. "Progressive Taxation and the Equal Sacrifice Principle." *Journal of Public Economics* 32 (2): 203–14.