Eichner, Thomas; Runkel, Marco

Working Paper

Homo Oeconomicus as the Homo Moralis' Party Pooper: Heterogeneous Morality in Public Good Games

CESifo Working Paper, No. 11231

Provided in Cooperation with:
Ifo Institute – Leibniz Institute for Economic Research at the University of Munich

Suggested Citation: Eichner, Thomas; Runkel, Marco (2024) : Homo Oeconomicus as the Homo Moralis' Party Pooper: Heterogeneous Morality in Public Good Games, CESifo Working Paper, No. 11231, CESifo GmbH, Munich

This Version is available at:
https://hdl.handle.net/10419/302716

CESifo WORKING PAPERS

11231
2024

July 2024

Homo Oeconomicus as the Homo Moralis' Party Pooper: Heterogenous Morality in Public Good Games

Thomas Eichner, Marco Runkel

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.cesifo.org/en/wp
An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.cesifo.org/en/wp

CESifo *Working Paper No. 11231*

Homo Oeconomicus as the Homo Moralis' Party Pooper: Heterogeneous Morality in Public Good Games

Abstract

The main insight of this paper is that moral behavior does not necessarily alleviate coordination problems or may even worsen them, if individuals possess different degrees of morality. We characterize heterogenous Alger-Weibull morality preferences in a canonical model of voluntary contributions to a public good. The analysis reveals a novel polarization effect which traces back to a 'preference for leadership' and weakens (strengthens) the incentive to contribute to the public good for individuals with below (above) average morality. Equilibrium public good provision is not increased by morality, as long as there are homo oeconomicus individuals. An increase in morality of an individual may reduce total provision of the public good, if heterogeneity is large enough. Redistributive transfers are no longer neutral.

JEL-Codes: C720, D910, H410.

Keywords: moral behaviour, Kantian ethics, heterogeneity, public goods.

Thomas Eichner
Department of Economics
University of Hagen
Universitätsstr. 41
Germany – 58097 Hagen
thomas.eichner@fernuni-hagen.de

Marco Runkel
Chair of Public Sector and Health Economics
Faculty of Economics and Management
University of Technology Berlin, H51
Straße des 17. Juni 135
Germany – 10623 Berlin
marco.runkel@tu-berlin.de

July 18, 2024
We would like to thank Jonas Kaiser as well as participants in research seminars at the Victoria University of Wellington and the University of Canterbury in Christchurch, in particular, Jeremy Clark, Eberhard Feess, Andrea Menclova and Vladimir Petkov, for their very useful comments. Financial support by the German Research Foundation (DFG) grant number EI 847/2-1 is gratefully acknowledged.

1 Introduction

Moral behavior is a widespread feature of human behavior (see, e.g., Bowles, 2017). It may be grounded by the categorical imperative introduced by the philosopher Immanuel Kant. Accordingly, the pure Kantian wants to do the right thing. She acts in accordance to that maxim through which she can at the same time want that it becomes a universal law (Kant, 1785). At first glance, one would intuitively expect that such a kind of moral behavior will mitigate coordination problems like the provision of public goods or the market failure due to externalities. The important contribution of the present paper is that this logic may not be true. Our main insight is that moral behavior does not necessarily alleviate coordination problems or may even worsen them, if individuals possess different degrees of morality. The basic intuition goes back to a novel polarization effect that provides individuals with a low degree of morality an incentive to counteract the ambitious attempts of individuals with high morality to mitigate the coordination problem.

This insight is brought forward within a framework of voluntary contributions to a pure public good, where individuals have *homo moralis* preferences tracing back to Alger and Weibull (2013, 2016, 2017, 2020). Accordingly, the *homo moralis* wants a certain fraction of the population to behave in the same way as she does. This fraction is interpreted as degree of morality and can formally be modeled by a (counterfactual) probability that other individuals behave as the *homo moralis* herself. The *homo moralis* then maximizes her corresponding expected material payoff. This quite general concept encompasses Kant's universalization principle not only in its pure form, according to which the *homo moralis* wants that all other individuals behave as she does, but also partial Kantian universalization, i.e. the *homo moralis* wants that only a fraction of other individuals chooses the same strategy as she does. The novel feature of our analysis is to extend the Alger-Weibull approach by allowing for *homo moralis* individuals with different degrees of morality.

Within this framework, we first derive a methodological result. Under the assumption that utility is non-linear in the public good, we characterize the marginal utility of the *homo moralis*. Heterogeneity in the degree of morality combined with non-linear utility makes this characterization technically challenging. To derive results, however, we apply generalizations of the so-called Chu-Vandermonde identity, which are known from combinatorics (see, e.g., Mestrovic 2018) but, to the best of our knowledge, have not yet been applied in economics. It then turns out that the marginal utility of the *homo moralis* contains the same terms as in the case with homogenous degrees of morality, i.e. a convex combination of the marginal utilities of the *homo oeconomicus*, who does not behave morally at all, and the *homo kantiensis*, who

behaves morally according to Kant's pure universalization principle (see, e.g., Alger and Weibull, 2016). But heterogeneity in morality augments the marginal utility of the *homo moralis* by a so far unknown polarization term that *ceteris paribus* weakens (strengthens) the incentive to contribute to the public good for individuals with below-(above-)average morality. Formally, we trace back the polarization incentive to a 'preference for leadership', i.e. the desire of the *homo moralis* to move public good provision away from the level preferred by others and closer to the level that she thinks is right.

Based on our new characterization of the *homo moralis* preferences within a heterogenous population, we re-examine the Nash equilibrium of the game of voluntary contributions to a public good and derive several unexpected insights. First, we consider utility functions that are quasi-linear in the private good as well as non-linear in the public good and derive a strong 'party pooper' result: the equilibrium public good provision is not increased at all by introducing morality, as long as there still is a group of *homo oeconomicus* individuals that have no moral preferences. Each attempt of moral individuals to improve public good provision is counteracted by *homo oeconomicus* individuals that reduce their contributions to the same extent as moral individuals increase theirs. Hence, *homo oeconomicus* individuals are the party poopers for those individuals which behave morally and want to render public good provision more efficient. Similar, we derive a weaker form of the party pooper property: the efficient public good provision is obtained only if all individuals are *homo kantiensis* and apply pure Kantian universalization. As long as there are *homo moralis* individuals practicing only partial universalization and, thus, having a degree of morality less than one, the economy does not achieve the efficient outcome. With numerical simulations we show that both the strong and the weak party pooper result still hold, if utility is non-linear in both the private good and the public good, even though the strong party pooper result is mitigated in the sense that the *homo oeconomicus* individuals crowd out a large part but not the whole increase in public good provision of moral individuals.

Second, and perhaps most important, we derive a further party pooper result by considering a comparative static increase in the morality of one group of individuals in an economy without *homo oeconomicus* individuals (to abstract from the strong party pooper result) and with at least some individuals practicing partial universalization only (to avoid that the economy has already attained efficiency). Increasing morality of a group that has already above average morality may then decrease public good provision and, thereby, move the economy to a less efficient equilibrium. The intuition is directly linked to the polarization incentive mentioned above. The total effect of the increasing morality can be decomposed into an effect caused by a first-order stochastic dominance (FSD) shift in morality and an

effect caused by a mean-preserving spread (MPS) in morality. The FSD shift increases aggregate public good provision, whereas the MPS reduces it. With homogeneity in morality, the FSD effect always dominates the MPS effect, leading to an unambiguous increase in public good provision. With heterogeneity in morality, in contrast, the polarization incentive changes the relative strength of the two effects such that the MPS effect may dominate the FSD effect, if heterogeneity is sufficiently large. The less moral individuals can then be viewed as party poopers for those individuals which would like to improve public good provision by becoming more moral. By means of numerical simulations, also this party pooper result is confirmed if utility is non-linear in both the public good and the private good.

Finally, we investigate the effect of heterogeneity in morality on the so-called neutrality property. Neutrality holds if redistributive income transfers among individuals do not change aggregate public good provision, since the increase in the voluntary contributions of the transfer-receiving individuals is just crowded out by the decrease in the voluntary contributions of the transfer-giving individuals. We show that neutrality applies also in our framework, if the individuals' utility functions are quasi-linear either in the private good or the public good or if individuals have the same degree of morality. But if utility is non-linear in both goods and if individuals possess heterogenous morality, the neutrality property breaks down. For such circumstances, we provide examples where an income transfer from low- to high-moral individuals reduces aggregate public good provision, since low-moral individuals are the party poopers and more than compensate the increase in contribution levels of high-moral individuals. Reversing this result, however, identifies a way how governments can improve public good provision: redistributing income from high- to low-moral individuals increases aggregate public good provision, since the low-moral individuals increases contribution levels by more than the high-moral individuals reduce theirs.

Overall, our paper is the first to derive a polarization incentive due to heterogenous Kant-Alger-Weibull *homo moralis* preferences. This incentive is absent in the presence of other types of social preferences[1] and may lead to adverse effects regarding the efficiency of voluntary public good provision. Therefore, morality may be a socially detrimental and not beneficial characteristic of human behavior. We want to emphasize, though, that we do not accuse Kant of having developed an antisocial norm for ethical behavior. Instead, we think that people may interpret the categorical imperative in two diametral different ways. In the first interpretation, probably intended by Kant himself, the individual acts

[1] In a supplementary appendix, we prove this assertion formally for public good economies with individuals having altruistic, fairness or social norm preferences.

socially and behaves in a way that is best for the society as a whole. Hence, in choosing the 'Kantian maxim' the individual takes into account that other individuals might want to consume a different quantity of the public good than it actually prefers. In contrast, the *homo morals* approach assumes that the individual takes solely its own preferences as the basis for selecting the Kantian maxim. It acts self-centeredly and antisocially, i.e. it has a preference for leadership and moves public good provision closer to that level that it thinks is right, ignoring that others might want to consume a different level. Of course, it is an empirical question which kind of interpretation of Kant's categorical imperative people apply and a philosophical task to figure out what Kant really meant. These questions are beyond the scope of our theoretical paper and left for future research. But polarization phenomena like those which we currently observe, for instance, in the heated debate on climate change or refugee migration may indicate anecdotal evidence that also the antisocial interpretation of the categorical imperative and, thus, the *homo moralis* framework is not unrealistic.

The economic literature provides an increasing number of studies modeling moral behavior based on Kant's categorical imperative. Our paper is related to that branch of the literature which formalizes the individual's selection of the Kantian maxim described in the categorical imperative.[2] Bilodeau and Gravel (2004), for example, model Kant's universalization principle by an equivalence relation that allows to identify strategies which are morally equivalent. Within a game of voluntary provision of a public good, the corresponding Kantian equilibrium, if it exists, turns out to be Pareto efficient. Roemer (2010) determines the Kantian maxim by the so-called Kantian optimization protocol, assuming that a strategy profile is a Kantian equilibrium if no individual would like all players to alter their strategies by the same factor. Such a Kantian equilibrium is also Pareto efficient. An extension of the Roemer approach is given in Roemer (2015) and applied by Grafton *et al.* (2017), Dizarlar and Karagözoğlu (2023) and Roemer and Silvestre (2023). In contrast to our framework, however, none of these studies of moral behavior obtains the polarization incentive and the possible detrimental effects of increasing the individuals' morality.

The present paper is closest to the line of literature that started with Laffont (1975). As the above-mentioned authors, he also focus on deriving the Kantian maxim. His idea is that

[2]The literature also provides modeling approaches which do not strictly focus on the explicit derivation of the Kantian maxim, but nevertheless formalize moral behavior according to Kant's categorical imperative. A simple and very promising example is the recent paper of Feess *et al.* (2023) who model morality by subtracting costs for morally questionable decisions from the individual's material payoff. They distinguish between consequentialists, who suffer from the morality costs only if their decision has consequences for others, and deontologists, who always incur morality costs when making a morally questionable decision.

a moral individual maximizes its payoff under the assumption that everybody behaves as it does. Such an individual is termed *homo kantiensis*. Very similar, Daube and Ulph (2016) and Eichner and Pethig (2022) assume moral agents that maximize a convex combination of a *homo oeconomicus* and a *homo kantiensis* payoff. This approach goes back to the concept of the *homo moralis* developed by Alger and Weibull (2013, 2016, 2017, 2020), that is also the starting point of our analysis. However, all these studies, including most works by Alger and Weibull, have been mainly assumed homogenous populations, ignoring heterogenous morality, which is the main innovation of our paper. Exceptions are Alger and Weibull (2017) and Alger and Laslier (2022). More specifically, Alger and Weibull (2017) investigate coordination games that are played among heterogenous individuals. They find that morality induces populations to select more efficient equilibria and point out that '. . . the most morally motivated individuals take the lead and are followed by less morally motivated individuals . . .' (Alger and Weibull 2017, p. 19).[3] Within a political economy setting with heterogenous voters, Alger and Laslier (2022) show that *homo moralis* preferences may help individuals to overcome coordination problems in elections. These conclusions stand in stark contrast to our main insight that, due to a preferences for leadership and the associated polarization, increasing morality may reduce efficiency. The reason is that we assume non-linear utility. With linear utility, as considered in Alger and Weibull (2017) and Alger and Laslier (2022), also in our model polarization and the detrimental effects of morality disappear.

Our paper is also related to the insights derived by the literature on the canonical model of voluntary contributions to the provision of public goods that dates back to Warr (1982), Roberts (1984), Bergstrom *et al.* (1986) and Bernheim (1986). This type of model has many applications in economics, for example, the studies on global public goods and charitable donations surveyed by Andreoni and Payne (2013) and Buchholz and Sandler (2021). An important feature of the voluntary contribution model is that aggregate public good provision is inefficiently low and, in an interior equilibrium, redistributive lump-sum transfers or governmental contributions to the public good do not change aggregate public good provision. This property is known as neutrality or crowding out. But the previous studies focus on *homo oeconomicus* individuals, whereas we show that with non-linear utility and heterogenous *homo moralis* individuals neutrality breaks down.[4]

[3]Notice that Alger and Weibull (2017) also show that morality may induce less efficient equilibria if, for example, repeated games are considered, which are not addressed in our paper. Bomze *et al.* (2021) and Juan-Bartroli and Karagözoğlu (2024) consider heterogenous Alger-Weibull morality in settings with incomplete information or bargaining, which are also different from our canonical public good game.

[4]Other kinds of social preferences may destroy neutrality or crowding out as well. For example, Nyborg and Rege (2003) re-examine the crowding-out result and find that it holds in the pure altruism model of

The remainder of the paper is organized as follows. In Section 2, we briefly present the version of the canonical public good game which is relevant for our analysis. For this model, Section 3 characterizes the *homo moralis* preferences à la Alger-Weibull, if individuals have different degrees of morality. In Section 4, we investigate the Nash equilibrium of the public good game and derive the party pooper results. Section 5 studies the neutrality of redistributive transfer schemes and Section 6 concludes.

2 Basic framework

The model. Consider an economy with $m \geq 2$ groups of individuals. The group indices are $j, k, \ell = 1, \ldots, m$. The number of individuals in group k is $n^k \geq 1$, so our modeling encompasses the specific case where group k contains only one single individual. The total number of individuals equals $n := \sum_{\ell=1}^{m} n^\ell$. Within each group, all individuals are identical. We focus on the case where the groups differ only in their members' degree of morality, which we introduce in the next section. The representative individual of group k consumes x^k units of a private good and G units of a pure public good. Utility of this individual is[5]

$$u^k = W\big(x^k\big) + V\big(G\big), \tag{1}$$

where the derivatives of the sub-utility functions V and W satisfy $W_x > 0 \geq W_{xx}$ and $V_G > 0 \geq V_{GG}$, with at least one of the second derivatives W_{xx} and V_{GG} being strictly negative. Each individual disposes of an exogenously given endowment ω of a third good, which is the same across groups. Both the private and the public good are produced in a one-to-one relation from the endowment. The resource constraint in the economy reads

$$\sum_{\ell=1}^{m} n^\ell x^\ell + G = n\omega. \tag{2}$$

According to (2), the sum of public and private consumption equals total endowment.

Social optimum. In order to determine the socially optimal solution in the above economy, consider a social planer who maximizes the weighted utilitarian social welfare function

$$\sum_{\ell=1}^{m} \alpha^\ell n^\ell \Big[W\big(x^\ell\big) + V\big(G\big)\Big], \tag{3}$$

Andreoni (1988), whereas it does not hold in the fairness model of Fehr and Schmidt (1999), the impure altruism models of Andreoni (1990) and Holländer (1990) and in the social norm model of Brekke *et al.* (2003). See also the analysis in the recent paper of Faias *et al.* (2020).

[5]We use the notational convention that lower-case letters represent variables or parameters. Upper-case letters are reserved to denote functions and subscripts attached to them indicate derivatives. The only exemption from this rule is the aggregate public good provision G. Boldface letters indicate vectors.

subject to the resource constraint (2). The group welfare weights α^ℓ for all ℓ satisfy $\sum_{\ell=1}^m \alpha^\ell n^\ell = 1$. By standard Lagrangian methods, we obtain the first-order conditions

$$\sum_{\ell=1}^{m} n^\ell \frac{V_G\big(G^o\big)}{W_x\big(x^{\ell o}\big)} = 1, \tag{4}$$

$$W_x\big(x^{ko}\big) = \frac{\mu}{\alpha^k}, \quad \text{for all } k, \tag{5}$$

where the superscript o indicates the socially optimal solution and μ is the Langrange multiplier associated with the resource constraint (2). Equation (4) represents the standard Samuleson rule for the efficient provision of a pure public good. The efficient allocation of the private good is determined by (5) and depends on the social planner's distributional preferences specified by the welfare weights α^k for all k.[6]

Business-as-usual. In the decentralized economy, each individual of group k contributes g^k units to the provision of the public good. Its private budget constraint reads

$$x^k + g^k = \omega. \tag{6}$$

The total amount of the public good is the sum of the individuals' contribution, i.e.

$$G = \sum_{\ell=1}^{m} n^\ell g^\ell. \tag{7}$$

As benchmark, we consider the business-as-usual (BAU) scenario in which all individuals are *homo oeconomici* and have no moral preferences. The representative individual of group k then chooses its own contribution level g^k in order to maximizes its utility (1) subject to the budget constraint (6) and the public good provision (7), taking as given the contribution levels $\bar{g}^k$ of the other group k individuals and g^ℓ of the individuals in each other group $\ell \neq k$. Using again standard Lagrangian methods, it is straightforward to show that the Nash equilibrium in the BAU scenario is characterized by the first-order conditions

$$\frac{V_G\big(G^b\big)}{W_x\big(x^{\ell b}\big)} = 1, \tag{8}$$

$$W_x\big(x^{kb}\big) = \theta^k, \quad \text{for all } k, \tag{9}$$

[6]If utility is quasi-linear in the private good, what is frequently assumed in the subsequent analysis, we have $W_x = 1$ and obtain a social optimum only if $\alpha^\ell = 1/n$ for all ℓ. The efficient allocation then determines only aggregate private consumption $\sum_{\ell=1}^m n^\ell x^{\ell o}$, but not the distribution of private consumption levels $\{x^{1o}, \ldots, x^{\ell o}\}$. Put differently, the social planner cannot address distributional issues in this case.

where the superscript b indicates the BAU scenario and θ^k is the Lagrange multiplier associated with the private budget (6). In contrast to the Samuelson rule (4) for the socially optimal solution, equation (8) shows that in the BAU scenario each individual's marginal rate of substitution between public and private consumption (instead of the sum of marginal rates of substitution) has to be equal to the marginal rate of transformation. Hence, the BAU is characterized by inefficient underprovision of the public good.

3 *Homo moralis* preferences

In order to model moral behavior, we follow the *homo moralis* approach of Alger and Weibull (2013, 2016, 2017, 2020) and extend it according to the following definition.[7]

Definition 1. *An individual of group k is a homo moralis with degree of morality $\kappa^k \in [0,1]$ if its objective function reads*

$$\mathbb{E}u^k := \mathbb{E}\left[W\big(\omega - g^k\big) + V\left[g^k + \big(n^k - 1\big)\tilde{g}^k + \sum_{\ell=1,\ell\neq k}^{m} n^\ell \tilde{g}^\ell\right]\right], \tag{10}$$

where $\tilde{\boldsymbol{g}} := \big(\tilde{g}^1, \ldots, \tilde{g}^m\big)$ is a random contribution profile of the other individuals, with each component being the considered individual's contribution g^k with probability κ^k and the actual contribution of another individual, i.e. $\bar{g}^k$ if the other individual is from group k and g^ℓ if the other individual is from group $\ell \neq k$, with probability $1-\kappa^k$.

According to this definition, a *homo moralis* individual from group k replaces with probability κ^k another individual's contribution with its own contribution g^k, while with probability $1-\kappa^k$ it keeps the other individual's contribution, i.e. $\bar{g}^k$ in case of another individual of group k and g^ℓ in case of another individual of group $\ell \neq k$. The definition thus follows the partial Kantian universalization principle mentioned in the Introduction, with the universalization principle in its pure form obtained for $\kappa^k = 1$. We denote (10) as *homo moralis* utility. Definition 1 is the same as Equation (2) in Alger and Weibull (2017) and Definition 1 in Alger and Laslier (2022), except that we allow for heterogeneity in the morality parameter

[7] A slightly more general formalization of this definition would consider expected utility $\mathbb{E}u_i^k$ of individual i from group k and replace the argument of W and V in (10) by, respectively, $\omega - g_i^k$ and $g_i^k + \sum_{z=1,z\neq i}^{n^k} g_z^k + \sum_{\ell=1,\ell\neq k}^{m}\sum_{z=1}^{n^\ell} g_z^\ell$, where g_z^ℓ is the contribution of individual z from group ℓ. For notational convenience, however, we save one index by already assuming that within each group $\ell \neq k$ all individuals choose the same contribution $g_z^\ell = g^\ell$. For group k we differentiate between the individual under consideration and other individuals, in order to be aware that the individual under consideration optimize only with respect to its own contribution g^k and not the contribution $\bar{g}^k$ chosen by the other individuals from its group.

κ^k. This implies that individuals from different groups usually choose different contribution levels with important consequences for the characterization of the *homo moralis* preferences.

In order to characterize the *homo moralis* utility, assume $q \in \{0, \ldots, n-1\}$ other individuals choose the same contribution g^k as the group k individual under consideration, while all other individuals choose their own contribution levels. This happens with probability

$$\left(\kappa^k\right)^q \left(1-\kappa^k\right)^{n-1-q}. \tag{11}$$

Suppose among the q individuals that choose the same contribution level, a number of $r^k \in \{0, \ldots, n^k - 1\}$ individuals are from group k and a number of $r^\ell \in \{0, \ldots, n^\ell\}$ individuals are from group $\ell \neq k$. For notational convenience, define the vector $\boldsymbol{r} := (r^1, \ldots, r^m)$ and let $S^k(q) := \left\{\boldsymbol{r} \middle| \sum_{\ell=1}^m r^\ell = q,\, 0 \leq r^k \leq n^k - 1,\, 0 \leq r^\ell \leq n^\ell \text{ for all } \ell \neq k\right\}$ be the set of all possible $\boldsymbol{r}$ whose elements sum up to q. For each $\boldsymbol{r} \in S^k(q)$, there are

$$N(\boldsymbol{r}) = \binom{n^1}{r^1} \ldots \binom{n^k - 1}{r^k} \ldots \binom{n^m}{r^m} \tag{12}$$

permutations. $N(\boldsymbol{r})$ gives the number of possibilities to draw $\boldsymbol{r}$ from the set of all individuals except for the group k individual under consideration, where r^k individuals are chosen from the remaining group k and r^ℓ with $\ell \neq k$ individuals are chosen from group ℓ. For any given $\boldsymbol{r} \in S^k(q)$, total public good provision is given by

$$G\left(\boldsymbol{r}, q, g^k, \bar{g}^k, \boldsymbol{g^{-k}}\right) := (q+1)g^k + \left(n^k - 1 - r^k\right)\bar{g}^k + \sum_{\ell=1, \ell\neq k}^{m} \left(n^\ell - r^\ell\right) g^\ell, \tag{13}$$

where, for notational convenience, we define $\boldsymbol{g^{-k}} := \left(g^1, \ldots, g^{k-1}, g^{k+1}, \ldots, g^m\right)$ as the vector of actual contribution levels in all groups except for group k. With the help of (11)–(13), the *homo moralis* utility defined in (10) can be rewritten as

$$\mathbb{E}u^k = W\left(\omega - g^k\right) + \sum_{q=0}^{n-1} \left(\kappa^k\right)^q \left(1-\kappa^k\right)^{n-1-q} \left\{ \sum_{\boldsymbol{r} \in S^k(q)} N(\boldsymbol{r})\, V\Big[G\left(\boldsymbol{r}, q, g^k, \bar{g}^k, \boldsymbol{g^{-k}}\right)\Big] \right\}, \tag{14}$$

and has the first- and second derivatives

$$\begin{aligned} \frac{\mathrm{d}\mathbb{E}u^k}{\mathrm{d}g^k} &= -W_x\left(\omega - g^k\right) + \sum_{q=0}^{n-1} \left(\kappa^k\right)^q \left(1-\kappa^k\right)^{n-1-q} (q+1) \times \\ &\qquad \times \left\{ \sum_{\boldsymbol{r} \in S^k(q)} N(\boldsymbol{r})\, V_G\Big[G\left(\boldsymbol{r}, q, g^k, \bar{g}^k, \boldsymbol{g^{-k}}\right)\Big] \right\}, \end{aligned} \tag{15}$$

$$\begin{aligned} \frac{\mathrm{d}^2\mathbb{E}u^k}{\left(\mathrm{d}g^k\right)^2} &= W_{xx}\left(\omega - g^k\right) + \sum_{q=0}^{n-1} \left(\kappa^k\right)^q \left(1-\kappa^k\right)^{n-1-q} (q+1)^2 \times \\ &\qquad \times \left\{ \sum_{\boldsymbol{r} \in S^k(q)} N(\boldsymbol{r})\, V_{GG}\Big[G\left(\boldsymbol{r}, q, g^k, \bar{g}^k, \boldsymbol{g^{-k}}\right)\Big] \right\}, \end{aligned} \tag{16}$$

since $\partial G\left(\boldsymbol{r}, q, g^k, \bar{g}^k, \boldsymbol{g^{-k}}\right)/\partial g^k = q + 1$ from differentiating (13). The second derivative in (16) is negative due to $W_{xx} < 0$ or $V_{GG} < 0$, ensuring that the expected utility (14) is concave. Hence, the second-order condition of the *homo moralis'* maximization problem will always be satisfied and, in the subsequent analysis, we will solely focus on the first derivatives (15) that determines the *homo moralis'* first-order condition.

For comparison purpose, we briefly replicate the homogenous case in our model. Assume all individuals have the same morality $\kappa^k = \kappa$ for all k. In the appendix, we prove

Lemma 1. *Suppose individuals are homogenous, i.e. $\kappa^k = \kappa$ for all k. Then, for all $\kappa \in [0, 1]$ the marginal utility (15) can be rewritten as*

$$\frac{\mathrm{d}\mathbb{E}u^k}{\mathrm{d}g^k} = -W_x\big(\omega - g^k\big) + (1-\kappa)\, V_G\left[g^k + (n^k - 1)\bar{g}^k + \sum_{\ell=1,\ell\neq k}^{m} n^\ell g^\ell\right] + n\kappa\, V_G\big(ng^k\big). \quad (17)$$

Not surprisingly, the expression in (17) is analogous to the first derivative of the *homo moralis* utility derived in the previous literature, see Alger and Weibull (2017, 2022), for instance. Alternatively to deriving it from (15), it can conveniently and equivalently be obtained by taking the derivative of the utility function

$$u^k = W\big(\omega - g^k\big) + (1-\kappa)\, V\left[g^k + (n^k - 1)\bar{g}^k + \sum_{\ell=1,\ell\neq k}^{m} n^\ell g^\ell\right] + \kappa\, V\big(ng^k\big). \quad (18)$$

Equation (18) is the convex combination of the utility of a *homo oeconomicus* individual, which has $\kappa = 0$ and no moral preferences at all, and the utility of a *homo kantiensis* individual, which is located at the other end of the morality range with $\kappa = 1$ and full morality. Under homogeneity, the *homo moralis* thus represents a broad concept of a moral person, encompassing the *homo oeconomicus* and the *homo kantiensis* as polar cases, but also allowing for intermediate degrees of morality simply measured by the parameter κ.

While Lemma 1 replicates the case of homogenous morality in our framework, we now turn to the novel case of heterogenous individuals that may differ in the degree of morality. Accordingly, from now on we consider an economy where κ^k is distributed over the interval $[0, 1]$ and group-specific. For this economy, we characterize the *homo moralis* preferences for each group k. We begin with the preferences of the two polar groups containing *homo oeconomicus* and *homo kantiensis* individuals. The appendix proves

Lemma 2. *Suppose individuals are heterogenous. Then,*

(i) the marginal utility (15) for $\kappa^k = 0$ is the same as the marginal utility (17) for $\kappa = 0$,

(*ii*) *the marginal utility* (*15*) *for* $\kappa^k = 1$ *is the same as the marginal utility* (*17*) *for* $\kappa = 1$.

Lemma 2 shows that in the presence of heterogenous individuals the marginal utility of both the *homo oeconomicus* and the *homo kantiensis* is the same as with homogenous individuals. The intuition is straightforward. The *homo oeconomicus* is not moral at all. Hence, it is clear that her preferences are independent of whether the other individuals are equally moral or not. The same argument applies to the *homo kantiensis*, since she always assumes that all other individuals behave in the same way as she does.

In contrast to the polar groups consisting of *homo oeconomicus* and the *homo kantiensis* individuals, the decision of the *homo moralis* with intermediate morality $\kappa^k \in]0,1[$ depends on whether the individuals are homogenous or not. Unfortunately, an analysis with a general functional form of the sub-utility function V is not tractable for $\kappa^k \in]0,1[$. However, we get results, if we approximate V by a second-order Taylor expansion.

Lemma 3. *Suppose individuals are heterogenous and*

$$V(G) = \gamma_0 + \gamma_1 G - \frac{\gamma_2}{2} G^2, \quad \textit{with} \quad \gamma_0 \gtreqless 0 \quad \textit{and} \quad \gamma_1, \gamma_2 \geq 0. \tag{19}$$

Then, for $\kappa^k \in [0,1]$ *the marginal utility* (*15*) *can be written as*

$$\begin{aligned} \frac{\mathrm{d}\mathbb{E}u^k}{\mathrm{d}g^k} = &-W_x\big(\omega - g^k\big) + \big(1-\kappa^k\big)\, V_G\left[g^k + (n^k-1)\bar{g}^k + \sum_{\ell=1,\ell\neq k}^{m} n^\ell g^\ell\right] + n\kappa^k\, V_G\big(ng^k\big) \\ &+ \kappa^k(1-\kappa^k)(n-2)(n-1)\gamma_2 \left\{ g^k - \frac{\big(n^k-1\big)\bar{g}^k + \sum_{\ell=1,\ell\neq k}^{m} n^\ell g^\ell}{n-1} \right\}. \end{aligned} \tag{20}$$

If individuals are heterogenous, the marginal utility of a *homo moralis* with an intermediate morality $\kappa^k \in]0,1[$ equals her marginal utility in case of homogenous individuals, compare the first line of (20) with (17), augmented by a correction term represented by the second line of (20). In order to interpret the correction term, it is useful to draw back the first derivative in (20) to a different objective function. Alternatively to deducting (20) from (15), it can equivalently be obtained from taking the first derivative of

$$\begin{aligned} u^k = &\, W\big(\omega - g^k\big) + \big(1-\kappa^k\big)\, V\left[g^k + (n^k-1)\bar{g}^k + \sum_{\ell=1,\ell\neq k}^{m} n^\ell g^\ell\right] + \kappa^k\, V\big(ng^k\big) \\ &+ \kappa^k(1-\kappa^k)\frac{(n-2)(n-1)}{2}\gamma_2 \left\{ g^k - \frac{\big(n^k-1\big)\bar{g}^k + \sum_{\ell=1,\ell\neq k}^{m} n^\ell g^\ell}{n-1} \right\}^2. \end{aligned} \tag{21}$$

With heterogenous individuals, the *homo moralis* maximizes a convex combination of the *homo oeconomicus* and *homo kantiensis* utility, see the first line in (21), plus an additional

term that is quadratic in the deviation of the *homo moralis* public good contribution from the average contribution of all other individuals, see the second line in (21). The additional term represents the preference for leadership that follows from the antisocial interpretation of Kant's categorical imperative discussed in the Introduction. It is non-negative and reflects an extra benefit the *homo moralis* obtains, if she tries to move public good provision away from the average level preferred by others towards the level that she prefers herself.

Based on the preference for leadership, the correction term in the second line of the marginal utility (20) can be interpreted as a polarization incentive. A *homo moralis* of group k whose morality is above the average morality would like to move the economy to an above-average quantity of the public good. For this *homo moralis*, the polarization effect is positive and *ceteris paribus* represents an incentive to increase her contribution level g^k. The opposite holds if the *homo moralis* of group k has morality that is below the average. She would like to move the economy to a below-average public good quantity, so her polarization effect is negative, giving an incentive to *ceteris paribus* reduce her contribution level g^k. Notice that the polarization effect represents an extra incentive, in addition to the well-known strategic substitution (free riding) incentive inherent in public good games. For instance, if a shock gives an individual the incentive to increase its contribution level, all other individuals reduce their contribution levels because of free riding. In our framework with heterogenous *homo moralis* individuals, this standard strategic substitution incentive is still present, but augmented by the polarization incentive due to heterogenous morality.

There are four cases in which the polarization incentive disappears. First, if all individuals are identical and choose the same contribution $g^k = \bar{g}^k = g^\ell = g$, then the second line of (20) becomes zero and the polarization incentive is not present. This is consistent with the result in Lemma 1. Second, in the polar cases of the *homo oeconomicus* and the *homo kantiensis* we have $\kappa^k = 0$ and $\kappa^k = 1$, respectively, implying that the polarization incentive in the second line of (20) vanishes, consistently with the general result in Lemma 2. Third, the polarization incentive is zero if there are only two individuals ($n = 2$). This is plausible, since individual k then faces the situation where either all or none, but not a part, of the other individuals chose the same contribution. Finally, the polarization incentive also vanishes if the public good utility is linear ($\gamma_2 = 0$), since then the marginal utility does no longer depend on the public good. In our view, however, all these four cases are rather non-generic. Usually, in most economies we have more than two individuals, which are not identical, do not all behave like a *homo oeconomicus* or a *homo kantiensis* and receive concave utility from the public good. Thus, the polarization incentive is important and we will now derive its implications for the equilibrium of the public good game.

4 Nash Equilibrium of the public good game

For notational convenience, we denote by $O := \left\{\ell \,\middle|\, \kappa^\ell = 0\right\}$ the set of groups consisting of *homo oeconomicus* individuals, by $K := \left\{\ell \,\middle|\, \kappa^\ell = 1\right\}$ the set of groups containing *homo kantiensis* individuals and by $M := \left\{\ell \,\middle|\, \kappa^\ell \in]0,1[\right\}$ the set of groups encompassing *homo moralis* individuals with intermediate morality. For the sake of simplicity, the latter set is simply denoted as the set of groups with *homo moralis* individuals, even though the concept of the *homo moralis* contains the *homo oeconomicus* and the *homo kantiensis* as polar cases.

4.1 Utility quasi-linear in private good

Let us start with a utility function (1) that is quasi-linear in the private good, i.e. $W_x = 1$. If not stated otherwise, V may take a general functional form, not necessarily the quadratic approximation in (19). Since $W_x = 1$ implies $W_{xx} = 0$, we merely have to assume $V_{GG} < 0$. The socially optimal level of the public good is given by (4),[8] which determines only the aggregate amount of the public good. Since individuals solely differ in morality which does not play a role for the social optimum, we assume that - when the efficient quantity of the public good has to be provided by the individuals - each individual makes the same contribution. With that assumption, efficient public good provision is determined by

$$nV_G(G^o) = 1, \qquad g^{ko} = \frac{G^o}{n} =: g^o, \quad \text{for all } k. \tag{22}$$

For $W_x = 1$, the BAU scenario characterized by (8) and (9) also determines only the total quantity of the public good, G^b, but not the individual contribution g^{kb}. Since morality does not play a role for the BAU either, we also assume that each individual makes the same individual contribution in the BAU. Condition (8) therefore yields[9]

$$V_G(G^b) = 1, \qquad g^{kb} = \frac{G^b}{n} =: g^b, \quad \text{for all } k. \tag{23}$$

Comparing (22) with (23) and recalling $V_{GG} < 0$, it is obvious that $G^b < G^o$ and $g^b < g^o$, i.e. both aggregate and individual public good provision in the BAU is inefficiently low.

The social optimum and the BAU are the benchmarks for the Nash equilibrium with heterogenous individuals. The latter follows from setting the marginal utility (15) equal to zero for all individuals. Within each group all individuals choose the same contribution levels, i.e. $g^k = \bar{g}^k$ in group k and g^ℓ in group $\ell \neq k$, while across groups the contribution levels may differ. Indicating Nash equilibrium values by a star, the appendix proves

[8] As mentioned in footnote 6, with $W_x = 1$ we must set $\alpha^k = 1/n$ for all k and obtain $\mu = 1/n$ from (5).

[9] For $W_x = 1$, condition (9) implies $\theta^k = 1$ for all k.

Proposition 1. *Suppose individuals are heterogenous,* $W_x = 1$ *and* $V_{GG} < 0$*. If* $O \neq \emptyset$*, then*

$$G^* = G^b, \qquad g^{\ell *} \begin{cases} < g^b, & \text{for all } \ell \in O, \\ > g^b, & \text{for all } \ell \in M \cup K. \end{cases} \tag{24}$$

Proposition 1 represents the strong version of the 'party pooper' result mentioned in the Introduction. Regardless of how many individuals have moral preferences and regardless of how large these individuals' degree of morality is, the equilibrium provision of the public good remains at its BAU level as long as there is still at least one group of *homo oeconomici* in the population. This is bad news for all moral individuals: In the presence of *homo oeconomici*, moral individuals cannot improve the public good provision by becoming more moral. The intuition of this insight is that the *homo oeconomicus* free rides to 100% on the moral individuals. She reduces her public good contribution in a one-to-one relation whenever another individual becomes more moral and contributes more. Hence, the public good contribution of *homo oeconomicus* individuals is below the BAU level g^b, whereas *homo moralis* and *homo kantiensis* individuals contribute more than g^b.[10]

We may derive a further bad news for moral individuals, if we focus on the other end of the morality range. In the appendix, we show

Proposition 2. *Suppose individuals are heterogenous,* $W_x = 1$ *and* $V_{GG} < 0$*. If* $O \cup M \neq \emptyset$*, then* $G^* < G^o$*. If* $O \cup M = \emptyset$ *and* $K \neq \emptyset$*, then* $G^* = G^o$*.*

Proposition 2 states that the efficient solution cannot be attained, unless all individuals are *homo kantienses*. But an economy with only *homo kansiensis* individuals hardly exists in reality. Proposition 2 therefore also presents a kind of party pooper result, even though in a weaker form: *Homo kantienses* would like to implement the social optimum, but *homo moralis* and *homo oeconomicus* individuals prevent this, since their incomplete morality gives them the incentive to reduce their contribution below the efficient level.

In order to further improve our understanding of the impact that heterogeneity in morality exerts on the equilibrium public good provision, we next focus on an economy in which there is no *homo oeconomicus* (avoiding the BAU result obtained in Proposition 1) and in which not all individuals are *homo kantienses* (avoiding the efficient provision of the public good according to Proposition 2). Formally, we assume $O = \emptyset$ and $M \neq \emptyset$. Since

[10]Notice that the presence of *homo kantiensis* individuals does not contradict this result. They choose their contribution such that $nV(ng^k) = 1$ which does not imply $G^* = G^o > G^b$, but simply determines the individual contribution level g^k of the *homo kantiensis*.

Lemma 3 characterizes the preferences of individuals from groups $\ell \in M$ only for a quadratic approximation of the utility function $V(G)$, the subsequent analysis assumes (19). With the help of the replacement function known from aggregative games, the appendix derives a closed form solution of the equilibrium aggregate public good provision, which is given by

$$G^* = \frac{\gamma_1}{\gamma_2} - \frac{1}{\gamma_2} \frac{\displaystyle\sum_{\ell=1}^{m} \frac{n^\ell}{n} \frac{1}{\kappa^\ell \left[2I\left(1-\kappa^\ell\right) + n(1-I+I\kappa^\ell)\right]}}{\displaystyle\sum_{\ell=1}^{m} \frac{n^\ell}{n} \frac{1-\kappa^\ell+n\kappa^\ell}{\kappa^\ell \left[2I\left(1-\kappa^\ell\right) + n(1-I+I\kappa^\ell)\right]}}. \tag{25}$$

In (25) we have introduced an indicator variable $I \in \{0,1\}$ in order to highlight the role of the polarization effect identified in the second line of (20). In our model, $I = 1$ holds, but by setting $I = 0$ we can figure out the role the polarization effect plays for our results.

We are particularly interested in the effect of a unilateral increase in one group's morality on the equilibrium provision of the public good. As shown in the appendix, differentiating G^* from equation (25) with respect to morality κ^k yields

$$\frac{\partial G^*}{\partial \kappa^k} = \frac{1}{\gamma_2 \left(\sum_{\ell=1}^{m} B^\ell\right)^2} \frac{(n-1)n^k}{n(\kappa^k)^2\left[2I\left(1-\kappa^k\right) + n\left(1-I+I\kappa^k\right)\right]^2} \times$$
$$\times \sum_{\ell=1}^{m} \frac{n^\ell}{n} \frac{n\kappa^\ell - I(n-2)\left[\kappa^\ell\left(1-2\kappa^k\right) + \left(\kappa^k\right)^2\right]}{\kappa^\ell\left[2I\left(1-\kappa^\ell\right) + n\left(1-I+I\kappa^\ell\right)\right]}, \tag{26}$$

where B^ℓ is the summand in the denominator of the second term on the RHS of (25). From equation (26), it is straightforward to derive that without the polarization effect ($I = 0$), we always would have $\partial G^*/\partial \kappa^k > 0$, i.e. a unilateral increase in morality would unambiguously increase public good provision. However, correctly taking into account polarization ($I = 1$) may reverse the result. To see this, set $I = 1$ and rewrite (26) as

$$\frac{\partial G^*}{\partial \kappa^k} = \frac{1}{\gamma_2 \left(\sum_{\ell=1}^{m} B^\ell\right)^2} \frac{(n-1)n^k}{n(\kappa^k)^2\left[2 + (n-2)\kappa^k\right]^2} \times$$
$$\times \sum_{\ell=1}^{m} \frac{n^\ell}{n} \frac{\kappa^\ell\left[2 + (n-2)\kappa^k\right] + (n-2)\kappa^k\left(\kappa^\ell - \kappa^k\right)}{\kappa^\ell\left[2 + (n-2)\kappa^\ell\right]}. \tag{27}$$

If $\kappa^k < \kappa^\ell$ for all $\ell \neq k$, then the sum in the second line of (27) is still positive, implying again $\mathrm{d}G^*/\mathrm{d}\kappa^k > 0$. In contrast, if $\kappa^k > \kappa^\ell$ for at least one $\ell \neq k$, then the sign of $\partial G/\partial \kappa^k$ is indeterminate and may become negative. In this case, an increase in morality of one group may have the counterintuitive implication that total public provision declines.

In order to present an example and to improve our understanding of this counterintuitive result, we consider the special case with two groups, i.e. group k and group j, where group k has higher morality than group j, i.e. $\kappa^k = \kappa + \varepsilon > \kappa = \kappa^j$ with $\kappa \in [0, 1[$ and $\varepsilon \in]0, 1 - \kappa[$. It turns out to be useful to decompose the whole effect of a unilateral increase in group k's morality on public good provision into the effect caused by a first-order stochastic dominance (FSD) shift in morality and the effect caused by a mean preserving spread (MPS) in morality. Formally, an increase in morality of both groups by the same amount, i.e. $\mathrm{d}\kappa^k = \mathrm{d}\kappa^j$, is an FSD shift. It increases the mean of morality, but leaves unchanged the variance of morality. Changes of morality satisfying $\mathrm{d}\kappa^j = -\frac{n^k}{n^j}\mathrm{d}\kappa^k$ impliy an MPS and increases the variance of morality, but leaves unchanged the mean of morality.[11] Denoting equilibrium public good provision as a function $G^*(\kappa^k, \kappa^j)$ and differentiating yields

$$\frac{\mathrm{d}G}{\mathrm{d}\kappa^k}\Big|_{\mathrm{FSD}} = \frac{\partial G^*}{\partial \kappa^k} + \frac{\partial G^*}{\partial \kappa^j}, \qquad \frac{\mathrm{d}G}{\mathrm{d}\kappa^k}\Big|_{\mathrm{MPS}} = \frac{\partial G^*}{\partial \kappa^k} - \frac{n^k}{n^j}\frac{\partial G^*}{\partial \kappa^j}. \tag{28}$$

With the help of these expressions, we can write

$$\frac{\partial G^*}{\partial \kappa^k} = \frac{n^k}{n}\frac{\mathrm{d}G^*}{\mathrm{d}\kappa^k}\Big|_{\mathrm{FSD}} + \frac{n^j}{n}\frac{\mathrm{d}G^*}{\mathrm{d}\kappa^k}\Big|_{\mathrm{MPS}}. \tag{29}$$

According to equation (29), the comparative static effect of a unilateral increase in group k's degree of morality on public good provision is the sum of an FSD effect and an MPS effect. This is plausible since an increase in κ^k raises both the mean of morality, represented by the FSD shift, and the variance of morality, reflected by the MPS. We will now show that the polarization effect from Section 3 is key for the relative importance of the two effects on the RHS of (29) and, therefore, for the overall sign of (29).

Consider first the hypothetical case in which the polarization effect would not be present, i.e. $I = 0$. In the appendix, we prove

Lemma 4. *Suppose individuals are heterogenous, $W_x = 1$, $V(G)$ satisfies (19), $O = \emptyset$ and $M \neq \emptyset$. If $I = 0$, $m = 2$ and $\kappa^k = \kappa + \varepsilon > \kappa = \kappa^j$ with $\kappa \in [0, 1[$ and $\varepsilon \in]0, 1 - \kappa[$, then*
(i) an FSD shift in morality increases G^*,
(ii) an MPS in morality decreases G^*,
(iii) a unilateral increase in morality κ^k increases G^*.

An FSD shift increases morality of each group and, thereby, improves total public good provision, as shown in part (i) of Lemma 4. For an MPS, in contrast, one group becomes more and the other group less moral. The group with the increased morality raises public

[11] The appendix gives a proof of the effects of the FSD shift and MPS on the mean and variance of morality.

good provision, whereas the group with the lowered morality reduces public good provision. The latter effect overcompensates the former, so overall the MPS leads to a reduction in public good provision, as proven by part (ii) of Lemma 4. As argued above, the unilateral increase in group k's morality combines an FSD shift and an MPS. Hence, according to part (i) and (ii) of Lemma 4 and equation (29), we obtain two countervailing effects on public good provision, if group k's morality unilaterally increases. However, as shown in part (iii) of Lemma 4, the increase in G^* due to the FSD shift overcompensates the decrease in G^* due to the MPS, so overall the unilateral increase in κ^k raises public good provision.

If the polarization effect is correctly taken into account, i.e. $I = 1$, the appendix proves

Proposition 3. *Suppose individuals are heterogenous, $W_x = 1$, $V(G)$ satisfies (19), $O = \emptyset$ and $M \neq \emptyset$. If $I = 1$, $m = 2$ and $\kappa^k = \kappa + \varepsilon > \kappa = \kappa^j$ with $\kappa \in [0, 1[$ and $\varepsilon \in]0, 1 - \kappa[$, then*
(i) an FSD shift in morality increases G^*,
(ii) an MPS in morality decreases G^*,
(iii) a unilateral increase in morality κ^k decreases (increases) G^* if ε is large (small).

According to parts (i) and (ii) of Proposition 3, even with the polarization effect there are still countervailing driving forces of a unilateral increase in group k's morality on public good provision. The FSD effect again improves public good provision, whereas the MPS effect lowers public good provision. However, compared to the hypothetical case without the polarization effect captured by Lemma 4, the relative sizes of the FSD and MPS effects may now be different. If asymmetries between the two groups are small, the FSD effect is still stronger than the MPS effect and total public good provision rises. But provided asymmetries are sufficiently large, polarization strengthens the MPS effect such that it overcompensates the FSD effect and total public good provision declines as a reaction of making the more moral group k even more moral, as proven by part (iii) of Proposition 3.[12]

In sum, with heterogenous morality, a unilateral increase of the morality of individuals which are already rather moral, relatively to other individuals, may have the counterintuitive effect of lowering total public good provision. This insight may also be interpreted as a (weak) party pooper result: If moral individuals become even more moral and would like to improve public good provision, their good intention is counteracted by the polarization effect of the less moral individuals, so that overall public good provision may decline.

[12]Notice that this logic also explains our conclusion from (27), that a unilateral increase in the morality of the least moral group cannot produce the counterintuitive result of reducing total public good provision. Such a unilateral change increases the mean of morality, but at the same time lowers the variance of morality, so the FSD and the MPS effects both point into the direction of increasing total public good provision.

4.2 Utility quasi-linear in public good

The strong and weak party pooper results derived in Proposition 1 and 3 cannot show up, if the utility function is quasi-linear in public consumption and strictly concave in private consumption. We can straightforwardly prove this assertion with the help of Lemma 3. Suppose $W_{xx} < 0$ and $V_G = 1$. The marginal utility of the *homo moralis* with $\kappa^k \in [0, 1]$ is then given by (20) for $\gamma_2 = 0$. Obviously, $\gamma_2 = 0$ implies that the polarization effect in the second line of (20) vanishes and the first-order condition for utility maximization becomes

$$-W_x(\omega - g^{k*}) + 1 - \kappa^k + n\kappa^k = 0. \tag{30}$$

Equation (30) implies that the individual contribution level of group k individuals, g^{k*}, is an increasing function of this group's degree of morality κ^k and independent of the morality of all other groups. Hence, equilibrium total public good provision $G^* = \sum_{\ell=1}^{m} n^\ell g^{\ell*}$ is never fixed to the BAU level, as long as there is one group with strictly positive degree of morality (the strong party pooper result of Proposition 1 cannot occur) and a unilateral increase in morality of one group will always increase total public good provision (the weak party pooper result obtain in Proposition 3(iii) cannot occur).

4.3 Non-quasi-linear utility

Perhaps most relevant is the case where the utility function is quasi-linear neither in private nor in public consumption. The effects of morality on public good provision is then a mix of the effects derived in the two previous subsections. In this subsection, however, we will use numerical analysis in order to show that, for non-linear utility, the party pooper property is still present, even though not in the strong version as derived in Proposition 1.

In the numerical examples, we assume the functional forms $V(G) = \gamma_0 + \gamma_1 G - \gamma_2 \frac{G^2}{2}$ and $W(x^k) = \alpha_0 + \alpha_1 x^k - \alpha_2 \frac{(x^k)^2}{2}$. Consider the two groups j and k with $\alpha_1 = \gamma_1 = 100$, $\alpha_2 = \gamma_2 = 1$, $n^k = n^j = 10$ and $\omega = 100$. We denote by $g^{k*}(\kappa^j, \kappa^k)$ the public good provision of group k and by $G^*(\kappa^j, \kappa^k)$ the aggregate public good provision when the morality of group j is κ^j and the morality of group k is κ^k. Figure 1 illustrates public good provision in four scenarios. The orange line indicates the social optimum, whereas the green line represents the BAU with $(\kappa^j, \kappa^k) = (0, 0)$. In the scenario represented by the red line group j has no morality ($\kappa^j = 0$), whereas in the scenario represented by the blue line group j has high morality ($\kappa^j = 0.9$). Comparing the blue and red line shows that decreasing group j's morality from $\kappa^j = 0.9$ to $\kappa^j = 0$ drastically reduces aggregate public good provision such that even large moralities of group k are not effective in improving total provision of the

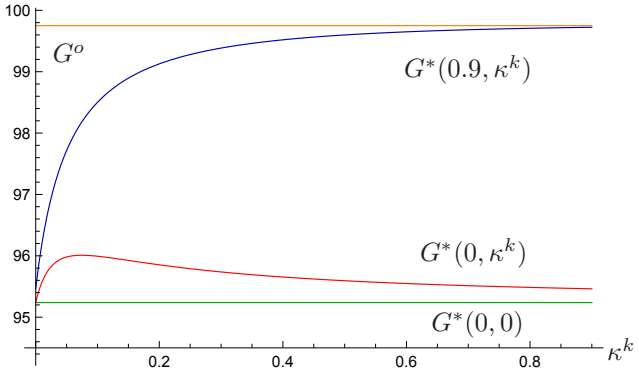

Figure 1: Public good provision in a numerical example with $\kappa^j \in \{0, 0.9\}$

public good. This insight is not the strong party pooper result in its pure form identified by Proposition 1, but it reveals that the party pooper behavior of less moral groups still may have a decisive impact on aggregate public good provision.

The next numerical example also relies on the above mentioned parameter values but now the morality of group j is fixed at $\kappa^j = 0.1$. In the left panel of Figure 2, the group with the higher degree of morality contributes more to the public good provision. The public

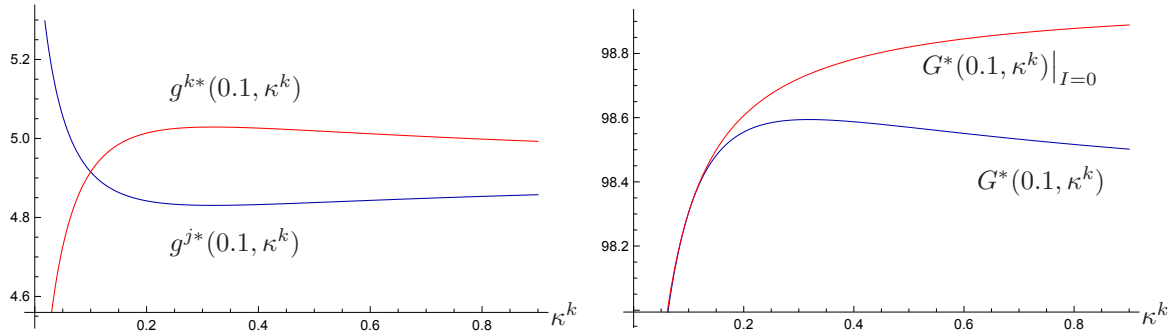

Figure 2: Public good provision in a numerical example with $\kappa^j = 0.1$

good provision of group k is increasing [decreasing] in κ^k if $\kappa^k < [>]0.319$. Conversely, public good provision of group j is decreasing [increasing] in κ^k if $\kappa^k < [>]0.319$. The blue line in the right panel of Figure 2 illustrates that the aggregate public good provision may decrease when the heterogeneity of the morality distribution raises, consistently with our party pooper result in Proposition 3 (this effect can also be seen at the red line in Figure 1, where, however, we fixed κ^j to zero, which was not done in Proposition 3). In view of the red line in the right panel of Figure 2, it becomes obvious that the preference for leadership and the corresponding polarization effect are again responsible for the party pooper result: when we ignore polarization ($I = 0$), total public good provision always increases as reaction on an increase of group k's morality and, thus, an increase of the heterogeneity of morality.

5 Neutrality

Heterogeneity in morality may also play a decisive role for the neutrality property of redistributive lump-sum transfers between the individuals. According to the neutrality property, redistributive lump-sum transfers change neither aggregate public good provision nor welfare. Formally, such a redistributive system consists of lump-sum payments τ^k from or to individuals of group k, where the transfers satisfy $\sum_{\ell=1}^{m} n^\ell \tau^\ell = 0$. The budget constraint of a group k individual turns into $x^k + g^k = \omega + \tau^k$. We may check the neutrality property within our framework by replacing ω by $\omega + \tau^k$ and considering a redistribute transfer increase (RTI) for group k individuals formally defined by $\mathrm{d}\tau^k = -\sum_{\ell=1,\ell\neq k}^{m} \frac{n^\ell}{n^k} \mathrm{d}\tau^\ell > 0$.

It is straightforward to prove that the RTI is neutral also in the presence of heterogenous morality, as long as utility is quasi-linear either in the private or the public good. In the former case, we have $W_x = 1$. Setting (15) equal to zero yields the first-order condition of utility maximization of a group k individual, which does not depend on $\omega + \tau^k$. Hence, individual public good provision g^k as well as total public good provision $G = \sum_{\ell=1}^{m} n^\ell g^\ell$ are not affected by an RTI. The same is true for the utilitarian welfare, since using $W_x = 1$ and $\sum_{\ell=1}^{m} n^\ell \tau^\ell = 0$ in (14) implies that $\sum_{\ell=1}^{m} \mathbb{E}u^\ell$ does not depend on τ^k. For utility quasi-linear in the private good neutrality therefore holds. If the utility function is quasi-linear in the public good, the first-order condition (30) becomes $g^k = \omega + \tau^k - W_x^{-1}\left(1 - \kappa^k + n\kappa^k\right)$. Hence, an RTI influences individual provision of the public good, indeed, but summing over all individuals and taking into account $\sum_{\ell=1}^{m} n^\ell \tau^\ell = 0$ reveals that total public good provision $G = \sum_{\ell=1}^{m} n^\ell g^\ell$ is again independent of the RTI. The private budget constraint implies $x^k = \omega + \tau^k - g^k = W_x^{-1}(1 - \kappa^k + n\kappa^k)$. Private consumption as well as welfare therefore do not dependent on the RTI, too, and redistributive transfers are again neutral.

In contrast, neutrality breaks down, if utility is linear neither in the private nor the public good. To see this, we use quadratic sub-utility functions $V(G)$ and $W(x^k)$. In the appendix, we derive the associated equilibrium public good provision G^* and prove

Proposition 4. *Suppose the subutility functions are* $V(G) = \gamma_0 + \gamma_1 G - \frac{\gamma_2}{2} G^2$ *with* $\gamma_0 \gtreqless 0$, $\gamma_1, \gamma_2 \geq 0$ *and* $W(x^k) = \alpha_0 + \alpha_1 x^k - \frac{\alpha_2}{2}\left(x^k\right)^2$ *with* $\alpha_0 \gtreqless 0$, $\alpha_1, \alpha_2 \geq 0$.
(i) Suppose individuals are homogenous. Then, an RTI leaves G^* *unchanged.*
(ii) Suppose individuals are heterogenous. Then, an RTI with $\mathrm{d}\tau^\ell < 0$ *for all* $\ell \neq k$ *decreases* G^* *if* $\kappa^k > \kappa^\ell$ *for all* $\ell \neq k$ *and increases* G^* *if* $\kappa^k < \kappa^\ell$ *for all* $\ell \neq k$.
(iii) Suppose individuals are heterogenous. Then, an RTI with $\mathrm{d}\tau^j < 0$ *and* $\mathrm{d}\tau^\ell = 0$ *for all* $\ell \neq j, k$ *decreases* G^* *if* $\kappa^k > \kappa^j$ *and increases* G^* *if* $\kappa^k < \kappa^j$.

According to part (i) of Proposition 4, an RTI does not change equilibrium provision of the public good even for non-linear utility, if individuals have the same degree of morality. In this case, the increase in group k's transfer τ^k increases public good provision of group k individuals, indeed, but due to homogeneity and the redistributive property of the RTI all other individuals change their individual contribution levels such that aggregate public good provision G^* remains unaltered. This is no longer true, if individuals are heterogenous. Part (ii) of Proposition 4 presents an example where the increase in group k's transfer is financed by a reduction in the transfer of all other groups $\ell \neq k$. The RTI then lowers total public good provision G^* as long as the transfer receiving group k is the group with the highest degree of morality, i.e. $\kappa^k > \kappa^\ell$ for all $\ell \neq k$. The reason is that the reduction in the contribution level of the other groups $\ell \neq k$ more than compensates group k's increase in public good contribution. This is also a kind a party pooper result: If the government aims at increasing public good provision by supporting the most moral group, less moral individuals adjust their public good provision such that aggregate public good provision declines. Conversely, however, the government may effectively increase aggregate public good provision by supporting the least moral group of individuals, i.e. $\kappa^k < \kappa^\ell$ for all $\ell \neq k$. The more moral groups then receive lower transfers, indeed, but their reduction in public good provision does not outweigh the increase of public good contributions by the least moral group. While part (ii) of Proposition 4 only holds for an increase in the transfer to the most or the least moral group, we can construct further examples with transfers from or to medium morality groups and similar implications. According to part (iii) of Proposition 4, for instance, if redistribution is changed solely between two of the m groups, then public good provision declines (increases), if the recipient group has higher (lower) morality than the group that finances the extended redistribution.

6 Conclusion

In the present paper, we investigate the impact of morality á Alger and Weibull (2013, 2016, 2017, 2020) when agents differ with respect to their degree of morality. In contrast to the case with homogenous morality, the *homo moralis* then maximizes an objective that contains not only the convex combination of the *homo oeconomicus*' and the *homo kantiensis*' utility, but also an additional utility that reflects the *homo moralis*' preference for leadership. Using this objective in a canonical model of voluntary public good provision, we find that (i) *homo oeconomicus* individuals may offset moral individuals' additional efforts in their public good provision such that aggregate public good provision remains unaffected compared to an

economy with only *homo oeconomici*, and (ii) an increase in an individual's morality may decrease aggregate public good provision presupposed the heterogeneity in moralities is large and the increase in the individual's morality further increases the heterogeneity. In both cases, the efforts of moral individuals to enhance the public good provision is nullified by less moral individuals. Hence, the less moral individuals are the party poopers for the more moral individuals. The driving force for the party pooper results is a polarization effect which stems from the preference for leadership and incentivizes moral individuals to strengthen or weaken their voluntary public good provision depending on whether their morality is above or below the average morality. The polarization effect is novel and unique in the literature.

In particular this polarization effect points to possible future research related to our paper. First, from a methodological point of view, it might be interesting to know whether and, if so, how our characterization of heterogenous *homo moralis* preferences, inclusive of the polarization incentive, changes when we go beyond the second-order Taylor approximation of the utility from public good consumption. To the best of our knowledge, the mathematical literature provides generalizations of the Chu-Vandermonde identity that restricts our analysis to quadratic utility functions only. If mathematical research progresses and provides even more generalizations of the Chu-Vandermonde identity, we may re-examine our analysis with a higher-order Taylor approximation of the utility function. Second, from an applied point of view, it might be interesting to use econometric or experimental methods in order to figure out the individuals' degree of morality and its distribution. This may be helpful in quantifying the polarization effect when the *homo moralis* approach is applied to specific coordination problems in practice. Such further tasks are important, but beyond the scope of the present paper and therefore left for future research.

Appendix

Proof of Lemma 1. With identical degrees of morality $\kappa^k = \kappa$ for all k, all individuals are the same and we assume that all individuals choose the same contribution level $g^k = \bar{g}^k = g^\ell =: g$. The public good provision in (13) can then be written as

$$G\big(\boldsymbol{r}, q, g^k, \bar{g}^k, \boldsymbol{g^{-k}}\big) = ng, \tag{31}$$

where we have used $\sum_{\ell=1}^{m} r^\ell = q$. Due to (31), in the homogenous case $G(\cdot)$ no longer depends on $\boldsymbol{r}$ and q. In equation (15), we can then factor out $V_G\big[G(\cdot)\big]$ and obtain

$$\frac{\mathrm{d}\mathbb{E}u^k}{\mathrm{d}g^k} = -W_x\big(\omega - g^k\big) + V_G\big[G(\cdot)\big] \sum_{q=0}^{n-1} (\kappa)^q (1-\kappa)^{n-1-q} (q+1) \left\{ \sum_{\boldsymbol{r} \in S^k(q)} N(\boldsymbol{r}) \right\}. \tag{32}$$

The summation term containing $N(\boldsymbol{r})$ can be simplified using the Chu-Vandermonde identity well-known from combinatorics. The multinominal generalization of this identity proven, for example, in Identity 2.5 of Mestrovic (2018) implies[13]

$$\sum_{\boldsymbol{r}\in S^k(q)} N(\boldsymbol{r}) = \binom{n-1}{q}. \tag{33}$$

Using this identity in (32) gives

$$\frac{\mathrm{d}\mathbb{E}u^k}{\mathrm{d}g^k} = -W_x\big(\omega - g^k\big) + V_G\big[G(\cdot)\big] \sum_{q=0}^{n-1} \binom{n-1}{q} (\kappa)^q (1-\kappa)^{n-1-q} (q+1). \tag{34}$$

Let Q be a random variable with realization q and binominal distribution with parameters $n-1$ and κ. The term $\binom{n-1}{q}(\kappa)^q(1-\kappa)^{n-1-q}$ is the probability function of Q. Hence, equation (34) can be rewritten as

$$\begin{aligned} \frac{\mathrm{d}\mathbb{E}u^k}{\mathrm{d}g^k} &= -W_x\big(\omega - g^k\big) + V_G\big[G(\cdot)\big]\mathbb{E}(Q+1) \\ &= -W_x\big(\omega - g^k\big) + (1-\kappa)V_G\big[G(\cdot)\big] + n\kappa V_G\big[G(\cdot)\big], \end{aligned} \tag{35}$$

where we have used $\mathbb{E}(Q+1) = 1-\kappa+n\kappa$. Taking into account, that due to homogeneity the public good quantity can be written in disaggregated form $G(\cdot) = g^k + \big(n^k-1\big)\bar{g}^k + \sum_{\ell=1,\ell\neq k}^m n^\ell g^\ell$ or aggregated form $G(\cdot) = ng$ gives (17) and completes the proof of Lemma 1.

Proof of Lemma 2. For the *homo oeconomicus* ($\kappa^k = 0$), we have $\big(\kappa^k\big)^q\big(1-\kappa^k\big)^{n-1-q} = 1$ if $q = 0$ and $\big(\kappa^k\big)^q\big(1-\kappa^k\big)^{n-1-q} = 0$ if $q \in \{1,\ldots,n-1\}$, so (15) simplifies to

$$\frac{\mathrm{d}\mathbb{E}u^k}{\mathrm{d}g^k} = -W_x\big(\omega - g^k\big) + \sum_{\boldsymbol{r}\in S^k(q)} N(\boldsymbol{r})\, V_G\Big[G\left(\boldsymbol{r}, 0, g^k, \bar{g}^k, \boldsymbol{g^{-k}}\right)\Big]. \tag{36}$$

For $q = 0$, the sole element of $S^k(q)$ is $\boldsymbol{r} = \boldsymbol{0}$ and we can ignore the summation sign in (36). Moreover, (12) and (13) imply $N(\boldsymbol{0}) = 1$ and $G(\boldsymbol{0}, 0, g^k, \bar{g}^k, \boldsymbol{g^{-k}}) = g^k + (n^k-1)\bar{g}^k + \sum_{\ell=1,\ell\neq k}^m n^\ell g^\ell$. Inserting into (36) gives the same derivative as we obtain by setting $\kappa^k = 0$ in (17). For the *homo kantiensis* ($\kappa^k = 1$), it holds $\big(\kappa^k\big)^q\big(1-\kappa^k\big)^{n-1-q} = 1$ if $q = n-1$ and $\big(\kappa^k\big)^q\big(1-\kappa^k\big)^{n-1-q} = 0$ if $q \in \{0,\ldots,n-2\}$, so (15) becomes

$$\frac{\mathrm{d}\mathbb{E}u^k}{\mathrm{d}g^k} = -W_x\big(\omega - g^k\big) + \sum_{\boldsymbol{r}\in S^k(q)} N(\boldsymbol{r})\, n\, V_G\Big[G\left(\boldsymbol{r}, n-1, g^k, \bar{g}^k, \boldsymbol{g^{-k}}\right)\Big]. \tag{37}$$

[13] Actually, Mestrovic (2018) proves this identity for any positive integer q only, whereas we need it also for $q = 0$. However, it is obvious that the identity (33) also holds for $q = 0$, since $q = 0$ implies $\boldsymbol{r} = \boldsymbol{0}$ and $N(\boldsymbol{r}) = N(\boldsymbol{0}) = 1$ due to (12), so both sides of (33) are equal to one.

For $q = n-1$, $S^k(q)$ contains $\boldsymbol{r} = (n^1, \ldots, n^k - 1, \ldots, n^m) =: \boldsymbol{n}$ as the sole element, implying that we can again ignore the summation sign in (37). From (12) and (13), we obtain $N(\boldsymbol{n}) = 1$ and $G(\boldsymbol{n}, n-1, g^k, \bar{g}^k, \boldsymbol{g^{-k}}) = ng^k$. Inserting into (37) gives the same first derivative that we obtain for the *homo kantiensis* by setting $\kappa = 1$ in (17).

Proof of Lemma 3. The key difference to the proof of Lemma 1 under homogeneity is that with heterogenous individuals we cannot safely assume that all individuals choose the same contribution level. In contrast to (31), the expression $G\big(\boldsymbol{r}, q, g^k, \bar{g}^k, \boldsymbol{g^{-k}}\big)$ then still depends on $\boldsymbol{r}$ and q and cannot be factored out in (15). Simplifications of (15) can only be obtained if the functional form of $V(G)$ is specified. If it satisfies (19), equation (15) reads

$$\begin{aligned}\frac{\mathrm{d}\mathbb{E}u^k}{\mathrm{d}g^k} &= -W_x\big(\omega - g^k\big) + \sum_{q=0}^{n-1}\big(\kappa^k\big)^q\big(1-\kappa^k\big)^{n-1-q}(q+1)\times \\ &\qquad \times\left\{\sum_{\boldsymbol{r}\in S^k(q)} N(\boldsymbol{r})\left[\gamma_1 - \gamma_2 G\left(\boldsymbol{r}, q, g^k, \bar{g}^k, \boldsymbol{g^{-k}}\right)\right]\right\}. \quad (38)\end{aligned}$$

Equation (13) can be rewritten as

$$G\left(\boldsymbol{r}, q, g^k, \bar{g}^k, \boldsymbol{g^{-k}}\right) := qg^k + \big(n^k - 1\big)\big(\bar{g}^k - g^k\big) + \sum_{\ell=1}^{m} n^\ell g^\ell - \sum_{\ell=1}^{m} r^\ell g^\ell - r^k\big(\bar{g}^k - g^k\big). \quad (39)$$

Inserting (39) into (38), we obtain

$$\frac{\mathrm{d}\mathbb{E}u^k}{\mathrm{d}g^k} = -W_x\big(\omega - g^k\big) + \Phi_1 - \Phi_2 - \Phi_3 - \Phi_4 + \Phi_5 + \Phi_6, \quad (40)$$

with

$$\Phi_1 := \gamma_1 \sum_{q=0}^{n-1}\big(\kappa^k\big)^q\big(1-\kappa^k\big)^{n-1-q}(q+1)\sum_{\boldsymbol{r}\in S^k(q)} N(\boldsymbol{r}), \quad (41)$$

$$\Phi_2 := \gamma_2\, g^k \sum_{q=0}^{n-1}\big(\kappa^k\big)^q\big(1-\kappa^k\big)^{n-1-q}(q+1)q\sum_{\boldsymbol{r}\in S^k(q)} N(\boldsymbol{r}), \quad (42)$$

$$\Phi_3 = \gamma_2\big(n^k - 1\big)\big(\bar{g}^k - g^k\big)\sum_{q=0}^{n-1}\big(\kappa^k\big)^q\big(1-\kappa^k\big)^{n-1-q}(q+1)\sum_{\boldsymbol{r}\in S^k(q)} N(\boldsymbol{r}), \quad (43)$$

$$\Phi_4 := \gamma_2\left(\sum_{\ell=1}^{m} n^\ell g^\ell\right)\sum_{q=0}^{n-1}\big(\kappa^k\big)^q\big(1-\kappa^k\big)^{n-1-q}(q+1)\sum_{\boldsymbol{r}\in S^k(q)} N(\boldsymbol{r}), \quad (44)$$

$$\Phi_5 := \gamma_2 \sum_{q=0}^{n-1}\big(\kappa^k\big)^q\big(1-\kappa^k\big)^{n-1-q}(q+1)\sum_{\boldsymbol{r}\in S^k(q)} N(\boldsymbol{r})\left(\sum_{\ell=1}^{m} r^\ell g^\ell\right), \quad (45)$$

$$\Phi_6 \;:=\; \gamma_2 \sum_{q=0}^{n-1} \big(\kappa^k\big)^q \big(1-\kappa^k\big)^{n-1-q}(q+1) \sum_{\boldsymbol{r}\in S^k(q)} N(\boldsymbol{r})\, r^k \left(\bar{g}^k - g^k\right). \tag{46}$$

The terms with $N(\boldsymbol{r})$ can again be simplified using the Chu-Vandermonde in (33). However, as in some of these terms $N(\boldsymbol{r})$ is multiplied by further terms, we need a further generalization of the identity. The generalization proven in Identity 2.1 of Mestrovic (2018) gives[14]

$$\sum_{\boldsymbol{r}\in S^k(q)} N(\boldsymbol{r}) \left(\sum_{\ell=1}^{m} r^\ell g^\ell \right) \;=\; \binom{n-1}{q} \cdot \frac{q\Big[\big(n^k-1\big)g^k + \sum_{\ell=1,\ell\neq k} n^\ell g^\ell\Big]}{n-1}. \tag{47}$$

$$\sum_{\boldsymbol{r}\in S^k(q)} N(\boldsymbol{r})\, r^k \left(\bar{g}^k - g^k\right) \;=\; \binom{n-1}{q} \cdot \frac{q\left(n^k-1\right)\left(\bar{g}^k - g^k\right)}{n-1}. \tag{48}$$

Using the identities (33), (47) and (48) in (41)–(46), we obtain

$$\Phi_1 \;=\; \gamma_1 \sum_{q=0}^{n-1} \binom{n-1}{q} \big(\kappa^k\big)^q \big(1-\kappa^k\big)^{n-1-q}(q+1), \tag{49}$$

$$\Phi_2 \;=\; \gamma_2\, g^k \sum_{q=0}^{n-1} \binom{n-1}{q} \big(\kappa^k\big)^q \big(1-\kappa^k\big)^{n-1-q}(q+1)q, \tag{50}$$

$$\Phi_3 \;=\; \gamma_2 \left(n^k-1\right)\left(\bar{g}^k - g^k\right) \sum_{q=0}^{n-1} \binom{n-1}{q} \big(\kappa^k\big)^q \big(1-\kappa^k\big)^{n-1-q}(q+1), \tag{51}$$

$$\Phi_4 \;=\; \gamma_2 \left(\sum_{\ell=1}^{m} n^\ell g^\ell \right) \sum_{q=0}^{n-1} \binom{n-1}{q} \big(\kappa^k\big)^q \big(1-\kappa^k\big)^{n-1-q}(q+1), \tag{52}$$

$$\Phi_5 \;=\; \gamma_2\, \frac{\big(n^k-1\big)g^k + \sum_{\ell=1,\ell\neq k} n^\ell g^\ell}{n-1} \sum_{q=0}^{n-1} \binom{n-1}{q} \big(\kappa^k\big)^q \big(1-\kappa^k\big)^{n-1-q}(q+1)q, \tag{53}$$

$$\Phi_6 \;=\; \gamma_2\, \frac{\big(n^k-1\big)\big(\bar{g}^k - g^k\big)}{n-1} \sum_{q=0}^{n-1} \binom{n-1}{q} \big(\kappa^k\big)^q \big(1-\kappa^k\big)^{n-1-q}(q+1)q. \tag{54}$$

Let Q_k be a random variable with realization q and binominal distribution with parameters $n-1$ and κ^k. The term $\binom{n-1}{q}\big(\kappa^k\big)^g\big(1-\kappa^k\big)^{n-1-q}$ is the probability function of Q^k and implies

$$\Phi_1 = \gamma_1\, \mathbb{E}\big(Q_k+1\big), \;\; \Phi_2 = \gamma_2\, g^k\, \mathbb{E}\big(Q_k^2+Q_k\big), \;\; \Phi_3 = \gamma_2 \left(n^k-1\right)\big(\bar{g}^k - g^k\big)\mathbb{E}\big(Q_k+1\big), \tag{55}$$

$$\Phi_4 = \gamma_2 \left(\sum_{\ell=1}^{m} n^\ell g^\ell \right) \mathbb{E}\big(Q_k+1\big), \quad \Phi_6 = \gamma_2\, \frac{\big(n^k-1\big)\big(\bar{g}^k - g^k\big)}{n-1} \mathbb{E}\big(Q_k^2+Q_k\big), \tag{56}$$

$$\Phi_5 = \gamma_2\, \frac{\big(n^k-1\big)g^k + \sum_{\ell=1,\ell\neq k} n^\ell g^\ell}{n-1}\, \mathbb{E}\big(Q_k^2+Q_k\big). \tag{57}$$

[14]Again, Mestrovic (2018) proves these identities only for $q>0$. But $q=0$ implies $\boldsymbol{r}=\boldsymbol{0}$ and $N(\boldsymbol{r}) = N(\boldsymbol{0}) = 1$. Inserting into (47) and (48) shows that both sides of the identities become zero.

Inserting (55)–(57) into (40) and collecting common terms, we obtain

$$\begin{aligned}\frac{\mathrm{d}\mathbb{E}u^k}{\mathrm{d}g^k} &= -W_x(\omega - g^\kappa) + \mathbb{E}(Q+1)\left\{\gamma_1 - \gamma_2\left[g^k + (n^k-1)\bar{g}^k + \sum_{\ell=1,\ell\neq k}^{m} n^\ell g^\ell\right]\right\} \\ &\quad - \mathbb{E}(Q^2+Q)\gamma_2\left[g^k - \frac{(n^k-1)\bar{g}^k + \sum_{\ell=1,\ell\neq k}^{m} n^\ell g^\ell}{n-1}\right], \end{aligned} \tag{58}$$

where the expected values can be computed as

$$\mathbb{E}\big(Q_k+1\big) = 1-\kappa^k+n\kappa^k, \quad \mathbb{E}\big(Q_k^2+Q_k\big) = (n-1)\kappa^k\big[2(1-\kappa^k)+n\kappa^k\big]. \tag{59}$$

Inserting (59) into (58) and further rearranging yields (20), which completes the proof.

Proof of Proposition 1. If $O \neq \emptyset$, then at least one group k consists of *homo oeconomicus* individuals with morality $\kappa^k = 0$. Following Lemma 2i, the first-order condition of utility maximization of individuals from this group can be derived by setting (17) equal to zero, taking into account $W_x = 1$, $\kappa = 0$ and $g^{k*} = \bar{g}^{k*}$. This results in

$$V\left(\sum_{\ell=1}^{m} n^\ell g^{\ell*}\right) = 1, \tag{60}$$

implying $G^* = \sum_{\ell=1}^{m} n^\ell g^{\ell*} = G^b$ from (23). Next consider the groups $k \in M \cup K$. All members of these groups have $\kappa^k > 0$ and the marginal utility given by equation (15). If all individuals would choose the individual contribution level from the BAU scenario, i.e. $g^\ell = g^b$ for all ℓ, equation (13) implies $G\left(\boldsymbol{r}, q, g^k, \bar{g}^k, \boldsymbol{g^{-k}}\right) = ng^b = G^b$ for all $\boldsymbol{r} \in S^k(q)$. Remember that $\sum_{\ell=1}^{m} r^\ell = q$ for all $\boldsymbol{r} \in S^k(q)$. Inserting this into the marginal utility of group k from (15) and using the Chu-Vandermonde identity (33) implies

$$\begin{aligned}\left.\frac{\mathrm{d}\mathbb{E}u^k}{\mathrm{d}g^k}\right|_{g^k=g^b, G(\ldots)=G^b} &= -1 + V\big(G^b\big)\sum_{q=0}^{n-1}\binom{n-1}{q}\big(\kappa^k\big)^q\big(1-\kappa^k\big)^{n-1-q}(q+1) \\ &= -1 + V\big(G^b\big)\mathbb{E}\big(Q_k+1\big) = (n-1)\kappa^k > 0, \end{aligned} \tag{61}$$

where in the last line we have used $V\big(G^b\big) = 1$ from (23) and $\mathbb{E}\big(Q_k+1\big) = 1-\kappa^k+n\kappa^k$ from (59). According to (61), at the BAU scenario, each individual from groups $k \in M \cup K$ has an incentive to increase its own contribution g^k and, thus, to set it above the BAU level, i.e. $g^{k*} > g^b$. Since the total quantity of the public good is fixed at G^b, it follows $g^{k*} < g^b$ for individuals from each group $k \in O$. This completes the proof of Proposition 1.

Proof of Proposition 2. The proof of the first part is by contradiction. Suppose $O \cup M \neq \emptyset$ implies $G^* = G^o$. Then, $g^{k*} = g^o$ for all k and $G\left(\boldsymbol{r}, q, g^k, \bar{g}^k, \boldsymbol{g^{-k}}\right) = ng^o = G^o$ for all $\boldsymbol{r} \in S^k(q)$. From the marginal utility (15) and the Chu-Vandermonde identity (33) we obtain

$$\begin{aligned} \left.\frac{\mathrm{d}\mathbb{E}u^k}{\mathrm{d}g^k}\right|_{g^k=g^o, G(\ldots)=G^o} &= -1 + V\big(G^o\big) \sum_{q=0}^{n-1} \binom{n-1}{q} \big(\kappa^k\big)^q \big(1-\kappa^k\big)^{n-1-q} (q+1) \\ &= -1 + V\big(G^o\big)\mathbb{E}\big(Q_k + 1\big) = -\frac{(n-1)(1-\kappa^k)}{n} < 0, \end{aligned} \tag{62}$$

where we have used (22) and (59). Hence, all individuals of groups $k \in K$ will stick to the contribution level g^o, since they have $\kappa^k = 1$, while all individuals from groups $k \in O \cup M$ have an incentive to reduce their contribution below g^o, due to $\kappa^k < 1$. It follows $G^* < G^o$, a contradiction. For proving the second part of Proposition 2, suppose $O \cup M = \emptyset$ and $K \neq \emptyset$. All individuals in the economy then have $\kappa^k = 1$ and choose the same contribution level g^*. Applying Lemma 2ii, equation (17) implies the equilibrium condition $-1 + nV_G(ng^*) = 0$ and $G^* := ng^* = G^o$ by (22). This completes the proof of Proposition 2.

Derivation of (25) and (26). For $O = \emptyset$ and $M \neq \emptyset$, neither is there a *homo oeconomicus* that fixes public good provision to the BAU level G^b nor are all individuals *homo kantienses* rendering public good provision equal to the social optimal level G^o. Assuming (19) implies that all individuals have the marginal utility (20). The first-order condition of utility maximization of an individual from group k is derived by setting this marginal utility equal to zero. In doing so, we multiply the second line of (20) by the indicator variable $I \in \{0, 1\}$ in order to isolate the impact that the polarization effect exerts on our results. Taking into account $W_x = 1$, $V_G(G) = \gamma_1 - \gamma_2 G$, $g^k = \bar{g}^k$, $G = \sum_{\ell=1}^m n^\ell g^\ell$ we get from (20)

$$\begin{aligned} (1-\kappa^k)\left(\gamma_1 - \gamma_2 \sum_{\ell=1}^m n^\ell g^\ell\right) + n\kappa^k\left(\gamma_1 - \gamma_2 n g^k\right) & \\ + I\kappa^k(1-\kappa^k)(n-2)\gamma_2\left(ng^k - \sum_{\ell=1}^m n^\ell g^\ell\right) = 1, & \end{aligned} \tag{63}$$

which can be rearranged to

$$\left(1 - \kappa^k + n\kappa^k\right)\left(\gamma_1 - \gamma_2 G\right) - \gamma_2\kappa^k\left[2I\big(1-\kappa^k\big) + n\big(1 - I + I\kappa^k\big)\right]\big(ng^k - G\big) = 1, \tag{64}$$

where $G = \sum_{\ell=1}^m n^\ell g^\ell$. Solving (64) with respect to ng^k yields

$$\begin{aligned} ng^k = G + \frac{1 - \kappa^k + n\kappa^k}{\kappa^k\left[2I\big(1-\kappa^k\big) + n\big(1 - I + I\kappa^k\big)\right]}\left(\frac{\gamma_1}{\gamma_2} - G\right) & \\ - \frac{1}{\gamma_2\kappa^k\left[2I\big(1-\kappa^k\big) + n\big(1 - I + I\kappa^k\big)\right]}. & \end{aligned} \tag{65}$$

Multiplying with n^k/n and rearranging, we obtain the expression

$$n^k g^k = \frac{n^k}{n}\left\{G + \frac{1-\kappa^k+n\kappa^k}{\kappa^k\left[2I\left(1-\kappa^k\right)+n\left(1-I+I\kappa^k\right)\right]}\left[\frac{\gamma_1}{\gamma_2} - G - \frac{1}{\gamma_2\left(1-\kappa^k+n\kappa^k\right)}\right]\right\}$$
$$=: R^k\left(G,\kappa^k\right). \tag{66}$$

$R^k\left(G,\kappa^k\right)$ is the so-called replacement function of group k (see Cornes and Hartley, 2007a, b). It is defined for all groups with $0 < \kappa^k \leq 1$, but not for groups with $\kappa^k = 0$, consistently with our assumption $O = \emptyset$. Summing the replacement function over all groups yields

$$\sum_{\ell=1}^{m} R^\ell\left(G,\kappa^\ell\right) = G. \tag{67}$$

Hence, the equilibrium provision of the public good, G^*, is the fixed-point of equation (67). Inserting $G = G^*$ and (66) into (67) yields

$$\sum_{\ell=1}^{m}\frac{n^\ell}{n}\left\{G^* + \frac{1-\kappa^\ell+n\kappa^\ell}{\kappa^\ell\left[2I\left(1-\kappa^\ell\right)+n\left(1-I+I\kappa^\ell\right)\right]}\left[\frac{\gamma_1}{\gamma_2} - G^* - \frac{1}{\gamma_2\left(1-\kappa^\ell+n\kappa^\ell\right)}\right]\right\} = G^*$$

or, equivalently,

$$\sum_{\ell=1}^{m}\frac{n^\ell}{n}\left\{\frac{1-\kappa^\ell+n\kappa^\ell}{\kappa^\ell\left[2I\left(1-\kappa^\ell\right)+n\left(1-I+I\kappa^\ell\right)\right]}\left[\frac{\gamma_1}{\gamma_2} - \frac{1}{\gamma_2\left(1-\kappa^\ell+n\kappa^\ell\right)}\right]\right\}$$
$$= G^*\sum_{\ell=1}^{m}\frac{n^\ell}{n}\frac{1-\kappa^\ell+n\kappa^\ell}{\kappa^\ell\left[2I\left(1-\kappa^\ell\right)+n\left(1-I+I\kappa^\ell\right)\right]}.$$

This equation can be solved with respect to

$$G^* = \frac{\displaystyle\sum_{\ell=1}^{m}\frac{n^\ell}{n}\left\{\frac{1-\kappa^\ell+n\kappa^\ell}{\kappa^\ell\left[2I\left(1-\kappa^\ell\right)+n\left(1-I+I\kappa^\ell\right)\right]}\left[\frac{\gamma_1}{\gamma_2} - \frac{1}{\gamma_2\left(1-\kappa^\ell+n\kappa^\ell\right)}\right]\right\}}{\displaystyle\sum_{\ell=1}^{m}\frac{n^\ell}{n}\frac{1-\kappa^\ell+n\kappa^\ell}{\kappa^\ell\left[2I\left(1-\kappa^\ell\right)+n\left(1-I+I\kappa^\ell\right)\right]}},$$

or, equivalenty, equation (25).

In order to prove (26), rewrite (25) as

$$G^* = \frac{\gamma_1}{\gamma_2} - \frac{1}{\gamma_2}\sum_{\ell=1}^{m} A^\ell \Big/ \sum_{\ell=1}^{m} B^\ell,$$

with

$$A^\ell := \frac{n^\ell}{n}\frac{1}{\kappa^\ell\left[2I\left(1-\kappa^\ell\right)+n\left(1-I+I\kappa^\ell\right)\right]},\ B^\ell := \frac{n^\ell}{n}\frac{1-\kappa^\ell+n\kappa^\ell}{\kappa^\ell\left[2I\left(1-\kappa^\ell\right)+n\left(1-I+I\kappa^\ell\right)\right]}.$$

Differentiating then yields

$$\frac{\partial G^*}{\partial \kappa^k} = -\frac{1}{\gamma_2}\frac{1}{\left(\sum_{\ell=1}^m B^\ell\right)^2}\left(A_k^k\sum_{\ell=1}^m B^\ell - B_k^k\sum_{\ell=1}^m A^\ell\right), \tag{68}$$

with the derivatives

$$\begin{aligned} A_k^k &= -\frac{n^k}{n}\frac{n - I(n-2)\big(1-2\kappa^k\big)}{\big(\kappa^k\big)^2\big[2I\big(1-\kappa^k\big)+n\big(1-I+I\kappa^k\big)\big]^2}, \\ B_k^k &= -\frac{n^k}{n}\frac{n - I(n-2)\big(1-2\kappa^k\big) + I(n-1)(n-2)\big(\kappa^k\big)^2}{\big(\kappa^k\big)^2\big[2I\big(1-\kappa^k\big)+n\big(1-I+I\kappa^k\big)\big]^2}. \end{aligned}$$

Inserting the derivatives into (68) and rearranging yields (26).

Effects of an FSD shift and an MPS on the mean and variance of morality. With only two groups k and j, the mean and variance of morality is defined as, respectively,

$$\mu = \frac{n^k}{n}\kappa^k + \frac{n^j}{n}\kappa^j, \qquad \sigma^2 = \frac{n^k}{n}\left(\kappa^k-\mu\right)^2 + \frac{n^j}{n}\left(\kappa^j-\mu\right)^2. \tag{69}$$

For the FSD shift, it holds $\mathrm{d}\kappa^k = \mathrm{d}\kappa^j$. Totally differentiating (69) then yields

$$\frac{\mathrm{d}\mu}{\mathrm{d}\kappa^k}\Big|_{\mathrm{FSD}} = \frac{\partial\mu}{\partial\kappa^k} + \frac{\partial\mu}{\partial\kappa^j} = 1, \quad \frac{\mathrm{d}\sigma^2}{\mathrm{d}\kappa^k}\Big|_{\mathrm{FSD}} = \frac{\partial\sigma^2}{\partial\kappa^k} + \frac{\partial\sigma^2}{\partial\kappa^j} + \frac{\partial\sigma^2}{\partial\mu}\frac{\mathrm{d}\mu}{\mathrm{d}\kappa^k}\Big|_{\mathrm{FSD}} = 0. \tag{70}$$

The MPS satisfies $\mathrm{d}\kappa^j = -\frac{n^k}{n^j}\mathrm{d}\kappa^k$ and it holds

$$\frac{\mathrm{d}\mu}{\mathrm{d}\kappa^k}\Big|_{\mathrm{MPS}} = \frac{\partial\mu}{\partial\kappa^k} - \frac{n^k}{n^j}\frac{\partial\mu}{\partial\kappa^j} = 0, \quad \frac{\mathrm{d}\sigma^2}{\mathrm{d}\kappa^k}\Big|_{\mathrm{MPS}} = \frac{\partial\sigma^2}{\partial\kappa^k} - \frac{n^k}{n^j}\frac{\partial\sigma^2}{\partial\kappa^j} = \frac{2n^k(\kappa^k-\kappa^j)}{n} > 0. \tag{71}$$

For a unilateral increase (ULI) in κ^k, we have $\mathrm{d}\kappa^k > 0 = \mathrm{d}\kappa^j$ and, thus,

$$\frac{\mathrm{d}\mu}{\mathrm{d}\kappa^k}\Big|_{\mathrm{ULI}} = \frac{\partial\mu}{\partial\kappa^k} = \frac{n^k}{n} > 0, \quad \frac{\mathrm{d}\sigma^2}{\mathrm{d}\kappa^k}\Big|_{\mathrm{ULI}} = \frac{\partial\sigma^2}{\partial\kappa^k} + \frac{\partial\sigma^2}{\partial\mu}\frac{\mathrm{d}\mu}{\mathrm{d}\kappa^k}\Big|_{\mathrm{ULI}} = \frac{2n^jn^k\left(\kappa^k-\kappa^j\right)}{n^2} > 0. \tag{72}$$

The ULI in κ^k therefore increases both the mean and variance of morality.

Proof of Lemma 4. For $I = 0$ and only two groups k and j, (26) simplifies to

$$\frac{\partial G^*}{\partial \kappa^k} = \frac{1}{\gamma_2(B^k+B^j)^2}\frac{n-1}{n^3}\frac{n^k}{(\kappa^k)^2} > 0, \tag{73}$$

which proves part (iii) of Lemma 4. Using (73) in (28) implies

$$\begin{aligned} \frac{\mathrm{d}G}{\mathrm{d}\kappa^k}\Big|_{\mathrm{FSD}} &= \frac{1}{\gamma_2(B^k+B^j)^2}\frac{n-1}{n^3}\left[\frac{n^k}{(\kappa^k)^2}+\frac{n^j}{(\kappa^j)^2}\right] > 0, &(74)\\ \frac{\mathrm{d}G}{\mathrm{d}\kappa^k}\Big|_{\mathrm{MPS}} &= -\frac{1}{\gamma_2(B^k+B^j)^2}\frac{(n-1)n^k}{n^3}\frac{(\kappa^k)^2-(\kappa^j)^2}{(\kappa^k)^2(\kappa^j)^2} < 0. &(75) \end{aligned}$$

which proves parts (i) and (ii) of Lemma 4.

Proof of Proposition 3. For $I = 1$ and only two groups k and j, (27) can be written as

$$\frac{\partial G}{\partial \kappa^k} = \theta \frac{n^k}{\left(\delta^k\right)^2} \left[n - n^j (n-2) \frac{\left(\kappa^j - \kappa^k\right)^2}{\delta^j} \right], \tag{76}$$

where we define

$$\theta := \frac{n-1}{\gamma_2 n^2 \left(\sum_{\ell=1}^m B^\ell \right)^2} > 0, \quad \delta^k := \kappa^k \left[2 + (n-2)\kappa^k \right] > 0. \tag{77}$$

Using $\kappa^k - \kappa^j = \varepsilon$, $\kappa^j = \kappa$ and δ^j according to (77) in (76) implies

$$\frac{\partial G}{\partial \kappa^k} \lesseqgtr 0 \qquad \Leftrightarrow \qquad \varepsilon \lesseqgtr \sqrt{\frac{n\kappa\left[2 + (n-2)\kappa\right]}{n^j (n-2)}}, \tag{78}$$

which proves part (iii) of Proposition 3. Making use of (76) in the first part of (28) yields

$$\frac{\mathrm{d}G^*}{\mathrm{d}\kappa}\Big|_{\mathrm{FSD}} = \frac{\theta}{\left(\delta^k \delta^j\right)^2} \Big[n n^k \left(\delta^j\right)^2 + n n^j \left(\delta^k\right)^2 - (n-2) n^k n^j \left(\delta^k + \delta^j\right) \left(\kappa^k - \kappa^j\right)^2 \Big].$$

Using the definition of δ^k and δ^j from (77), $\kappa^k = \kappa + \varepsilon$, $\kappa^j = \kappa$, $n^k = n - n^j$ and finally collecting terms with respect to ε, we obtain

$$\begin{aligned}
\frac{\mathrm{d}G^*}{\mathrm{d}\kappa}\Big|_{\mathrm{FSD}} \;=\; & \frac{\theta}{\left(\delta^k \delta^j\right)^2} \Big[n^2 \kappa^2 [2 + (n-2)\kappa]^2 + 4 n n^j \kappa \big[2 + 3(n-2)\kappa + (n-2)^2 \kappa^2 \big] \varepsilon \\
& + 2 n^j \big[2n + 2(n-2)\left(2n + n^j\right)\kappa + (n-2)^2 \left(2n + n^j\right)\kappa^2 \big] \varepsilon^2 \\
& + 2(n-2) n^j \left(n + n^j\right) [1 + (n-2)\kappa] \varepsilon^3 + (n-2)^2 \left(n^j\right)^2 \varepsilon^4 \Big] > 0,
\end{aligned}$$

which proves Proposition 3 (i). Inserting (76) into the second part of (28) implies

$$\frac{\mathrm{d}G^*}{\mathrm{d}\kappa^k}\Big|_{\mathrm{MPS}} = \frac{\theta n^k}{\left(\delta^k \delta^j\right)^2} \Big[n \Big[\left(\delta^j\right)^2 - \left(\delta^k\right)^2 \Big] - (n-2)\left(n^j \delta^j - n^k \delta^k\right) \left(\kappa^k - \kappa^j\right)^2 \Big].$$

Finally, employing again (77), $\kappa^k = \kappa + \varepsilon$, $\kappa^j = \kappa$, $n^k = n - n^j$ and collecting terms yields

$$\begin{aligned}
\frac{\mathrm{d}G^*}{\mathrm{d}\kappa^k}\Big|_{\mathrm{MPS}} \;=\; & -\frac{\theta n^k}{\left(\delta^k \delta^j\right)^2} \Big[4n\kappa \big[2 + 3(n-2)\kappa + (n-2)^2 \kappa^2 \big] \varepsilon \\
& + \big[4n + 2(n-2)\left(5n + 2n^j\right)\kappa + (n-2)^2 \left(5n + 2n^j\right)\kappa^2 \big] \varepsilon^2 \\
& + 2(n-2)\left(n + n^j\right) [1 + (n-2)\kappa] \varepsilon^3 + n^j (n-2)^2 \varepsilon^4 \Big] < 0,
\end{aligned}$$

which shows part (ii) and completes the proof of Proposition 3.

Proof of Proposition 4. For quadratic utility, the first-order condition of utility maximization of a group k individual is obtained by setting (20) equal to zero. Using the quadratic specification of $V(G)$ and $W(x^k)$, after some rearrangement we obtain

$$\alpha_2(\omega+\tau^k-g^k)+(1-\kappa^k+n\kappa^k)\left(\gamma_1-\gamma_2 G\right)-\gamma_2\kappa^k\left[2+(n-2)\kappa^k\right]\left(ng^k-G\right)=\alpha_1, \quad (79)$$

where $G=\sum_{\ell=1}^m n^\ell g^\ell$. Similar to (63)–(66), we solve (79) for g^k and multiply it with n^k in order to obtain the replacement function of group k, now denoted by $\widetilde{R}^k(G,\kappa^k)$. Solving for the fix point of $\sum_{\ell=1}^m \widetilde{R}^k(G^*,\kappa^k)=G^*$, the equilibrium public good provision reads

$$G^*=\frac{\alpha_2}{\phi}\sum_{\ell=1}^m\frac{n^\ell}{n}\frac{\omega+\tau^\ell}{\alpha_2/n+\gamma_2\delta^\ell}+\frac{1}{\phi}\sum_{\ell=1}^m\frac{n^\ell}{n}\frac{\gamma_1(1-\kappa^\ell+n\kappa^\ell)-\alpha_1}{\alpha_2/n+\gamma_2\delta^\ell}, \quad (80)$$

where, for notational convenience, we have used δ^k defined in (77) as well as

$$\phi:=\sum_{\ell=1}^m\frac{n^\ell}{n}\frac{\alpha_2/n+\gamma_2(1-\kappa^\ell+n\kappa^\ell)}{\alpha_2/n+\gamma_2\delta^\ell}>0. \quad (81)$$

Totally differentiating (80) yields

$$\mathrm{d}G^*=\frac{\alpha_2}{\phi}\sum_{\ell=1}^m\frac{n^\ell}{n}\frac{\mathrm{d}\tau^\ell}{\alpha_2/n+\gamma_2\delta^\ell}. \quad (82)$$

The RTI is defined by

$$\mathrm{d}\tau^k=-\sum_{\ell=1,\ell\neq k}\frac{n^\ell}{n^k}\mathrm{d}\tau^\ell>0. \quad (83)$$

If all groups are homogenous, then $\kappa^\ell=\kappa$, $\delta^\ell=\delta$ and $\alpha_2/n+\gamma_2\delta^\ell=\alpha_2/n+\gamma_2\delta$ for all ℓ. Inserting together with (83) in (82) implies $\mathrm{d}G^*=0$, which proves part (i) of Proposition 4.

In order to prove parts (ii) und (iii) of Proposition 4, rewrite (82) as

$$\mathrm{d}G^*=\frac{\alpha_2}{\phi}\left\{\frac{n^k}{n}\frac{\mathrm{d}\tau^k}{\alpha_2/n+\gamma_2\delta^k}+\sum_{\ell=1,\ell\neq k}^m\frac{n^\ell}{n}\frac{\mathrm{d}\tau^\ell}{\alpha_2/n+\gamma_2\delta^\ell}\right\}. \quad (84)$$

Using (83) in order to replace $\mathrm{d}\tau^k$, we obtain after some rearrangements

$$\mathrm{d}G^*=\frac{\alpha_2}{\phi}\sum_{\ell=1,\ell\neq k}^m\frac{n^\ell}{n}\mathrm{d}\tau^\ell\frac{\gamma_2(\delta^k-\delta^\ell)}{(\alpha_2/n+\gamma_2\delta^k)(\alpha_2/n+\gamma_2\delta^\ell)}. \quad (85)$$

If the RTI is such that $\mathrm{d}\tau^\ell<0$ for all $\ell\neq k$, then (85) yields $\mathrm{d}G^*<0$ if $\delta^k>\delta^\ell$ for all $\ell\neq k$. Since (77) implies that δ^k is positive and increasing in κ^k, we have $\mathrm{d}G^*<0$ if $\kappa^k>\kappa^\ell$ for all $\ell\neq k$. Similar, if $\kappa^k<\kappa^\ell$ for all $\ell\neq k$, then $\mathrm{d}G^*>0$, which completes the proof

of part (ii). If the RTI is such that $\mathrm{d}\tau^j < 0$ and $\mathrm{d}\tau^\ell = 0$ for all $\ell \neq j, k$, then (83) implies $\mathrm{d}\tau^k = -\frac{n^j}{n^k}\mathrm{d}\tau^j > 0$. Using in equation (84) gives

$$\mathrm{d}G^* = \frac{\alpha_2}{\phi}\frac{n^\ell}{n}\mathrm{d}\tau^\ell \frac{\gamma_2(\delta^k - \delta^j)}{(\alpha_2/n + \gamma_2\delta^k)(\alpha_2/n + \gamma_2\delta^j)}. \tag{86}$$

Hence, $\kappa^k > \kappa^j$ implies $\delta^k > \delta^j$ and $\mathrm{d}G^* < 0$. Conversely, for $\kappa^k < \kappa^j$ we obtain $\delta^k < \delta^j$ and $\mathrm{d}G^* > 0$, which completes the proof of part (iii) of Proposition 4.

References

Alger, I. and J.-F. Laslier (2022): Homo moralis goes to the voting booth: coordination and information aggregation. *Journal of Theoretical Politics* 34, 280-312.

Alger, I. and J.W. Weibull (2013): Homo moralis - preference evolution under incomplete information and assortative matching, *Econometrica* 81, 2269-2302.

Alger, I. and J.W. Weibull (2016): Evolution and Kantian morality, *Games and Economic Behavior* 98, 56-67.

Alger, I. and J.W. Weibull (2017): Strategic behavior of moralists and altruists, *Games* 8, 38.

Alger, I. and J. W. Weibull (2020): Morality: evolutionary foundations and economic implications, in Basu, K., Rosenblatt, D. and C. Sepulveda (eds.), *The State of Economics, the State of the World*, Cambridge, MIT Press.

Andreoni, J. (1988): Privately provided public goods in a large economy: The limits of altruism, *Journal of Public Economics* 35, 57-73.

Andreoni, J. (1990): Impure altruism and donations to public goods: A theory of warm-glow giving, *The Economic Journal* 100: 464-477.

Andreoni, J. and A.A. Payne (2013): Charitable giving, in Auerbach A.J., Chetty, R. Feldstein, M. and E. Saez (Eds.): *Handbook of Public Economics* 5, 1-50, Elsevier.

Bergstrom, T., Blume, L. and H. Varian (1986): On the private provision of public goods, *Journal of Public Economics* 29, 25-49.

Bernheim, B.D. (1986): On the voluntary and involuntary provision of public goods, *American Economic Review* 76, 787-793.

Bilodeau, M. and N. Gravel (2004): Voluntary provision of a public good and individual morality, *Journal of Public Economics* 88, 645-666.

Bomze, I., Schachinger, W. and J. Weibull (2021): Does moral play equilibriate?, *Economic Theory* 71, 305-315.

Bowles, S. (2017), *The moral economy: why good incentives are no substitute for good citizens*, Yale University Press.

Brekke, K.A., Kverndokk, S. and K. Nyborg (2003): An economic model of moral motivation, *Journal of Public Economics* 87, 1967-1983.

Buchholz, W. and T. Sandler (2021): Global public goods: A survey, *Journal of Economic Literature* 59, 488-545.

Cornes, R. and R. Hartley (2007a): Aggregative public good games, *Journal of Public Economic Theory* 9, 201-219.

Cornes, R. and R. Hartley (2007b): Weak links, good shots and other public good games: building on BBV, *Journal of Public Economics* 91, 1684-1707.

Daube, M. and D. Ulph (2016): Moral behaviour, altruism and environmental policy, *Environmental and Resource Economics* 63, 505-522.

Dizarlar, A. and E. Karagözoğlu (2023): Kantian equilibria of a class of Nash bargaining games, *Journal of Public Economic Theory* 25, 867-891.

Eichner, T. and R. Pethig (2022): Kantians defy the economists' mantra of uniform Pigouvian emissions taxes, *Ecological Economics* 200, 107514.

Faias M., Moreno-Garcia, E. and G.D. Myles (2020): Bergstrom, Blume, and Varian: Voluntary contributions and neutrality, *Journal of Public Economic Theory* 22, 285-301.

Feess, E., Kerzenmacher, F. and G. Muehlheusser (2023): Morally questionable decisions by groups: Guilt sharing and its underlying motives, *Games and Economic Behavior* 140, 380-400.

Fehr, E. and K.M. Schmidt (1999): A theory of fairness, competition, and cooperation, *Quarterly Journal of Economics* 114, 817-868.

Grafton, R.Q., Kompas, T. and N. van Long (2017): A brave new world? Kantian-Nashian interaction and the dynamics of global climate change mitigation, *European Economic*

Review 99, 31-42.

Holländer, H. (1990): A social exchange approach to voluntary cooperation. *American Economic Review* 80, 1157-1167.

Juan-Bartroli, P. and E. Karagözoğlu (2024): Moral preferences in bargaining, *Economic Theory*, forthcoming, https://doi.org/10.1007/s00199-023-01544-7.

Kant, I. (1785): *Grundlegung zur Metaphysik der Sitten* (1964, Groundwork of the Metaphysics of Morals, Harper Torchbooks, New York).

Laffont, J.-J. (1975): Macroeconomic constraints, economic efficiency and ethics: An introduction to Kantian economics, *Economica* 42, 430-437.

Mestrovic, R. (2018): Several generalizations and variations of Chu-Vandermonde identity, arXiv preprint arXiv:1807.10604.

Nyborg, K. and M. Rege (2003): Does public policy crowd out private contributions to public goods?, *Public Choice* 115, 397-418.

Roberts, R.D. (1984): A positive model of private charity and public transfers, *Journal of Political Economy* 92, 136-148.

Roemer, J.E. (2010): Kantian equilibrium, *Scandinavian Journal of Economics* 112, 1-24.

Roemer, J.E. (2015): Kantian optimization. A microfoundation for cooperation, *Journal of Public Economics* 127, 45-57.

Roemer, J.E. and J. Silvestre (2023): Kant and Lindahl, *The Scandinavian Journal of Economics*, in press.

Warr, P.G. (1982): Pareto optimal redistribution and private charity, *Journal of Public Economics* 19, 131-138.

Supplementary Appendix

In this supplementary appendix, we show that the detrimental effect of increasing morality caused by the polarization incentive cannot occur, if we consider other types of social preferences within our public good game. We will start with altruistic preferences, then turn to fairness preferences and finally consider a social norm.

Altruism. With altruistic preferences, the utility of an individual from group k reads

$$\begin{aligned} u^k &= W\big(\omega - g^k\big) + V\Big[g^k + (n^k-1)\bar{g}^k + \sum_{\ell=1,\ell\neq k}^{m} n^\ell g^\ell\Big] \\ &\quad + \kappa^k\Bigg\{(n^k-1)\Big[W\big(\omega-\bar{g}^k\big) + V\Big[g^k + (n^k-1)\bar{g}^k + \sum_{\ell=1,\ell\neq k}^{m} n^\ell g^\ell\Big]\Big] \\ &\quad + \sum_{\ell=1,\ell\neq k}^{m} n^\ell\Big[W\big(\omega-g^\ell\big) + V\Big[g^k + (n^k-1)\bar{g}^k + \sum_{j=1,j\neq k}^{m} n^j g^j\Big]\Big]\Bigg\}. \end{aligned} \tag{S1}$$

where $\kappa^k > 0$ is now the degree of altruism with which a group k individual takes into account the utility of the other group k individuals and of all individuals from the other groups $\ell \neq k$. Taking the derivatives of (S1) with respect to g^k yields

$$\frac{\mathrm{d}u^k}{\mathrm{d}g^k} = -W_x\big(\omega - g^k\big) + \big[1 + \kappa^k(n-1)\big]\, V_G(G), \tag{S2}$$

$$\frac{\mathrm{d}^2u^k}{\big(\mathrm{d}g^k\big)^2} = W_{xx}\big(\omega - g^k\big) + \big[1 + \kappa^k(n-1)\big]\, V_{GG}(G), \tag{S3}$$

where $G = \sum_{\ell=1}^{m} n^\ell g^\ell$ and where it has already been assumed that all group k individuals choose the same contribution level g^k. From (S3), we have $\mathrm{d}^2u^k/\big(\mathrm{d}g^k\big)^2 < 0$ due to $W_{xx} < 0$ and/or $V_{GG} < 0$, so the objective of a group k individual is always concave and we do not have to worry about second-order conditions of utility maximization.

In order to compare with Proposition 3, let us first assume quasi-linear utility in the private good, i.e. $W_x = 1$. In the Nash equilibrium, indicated by a star, we then have $g^{k*} > 0$ for group k with $\kappa^k = \max\{\kappa^1, \ldots, \kappa^m\}$ and $g^{\ell*} = 0$ for all $\ell \neq k$, so only the most altruistic group contributes to the public good and $G^* = n^k g^{k*}$. To see this, suppose the opposite, i.e. in equilibrium there is an $\bar{\ell} \neq k$ with $g^{\bar{\ell}*} > 0$. Using (S2) implies

$$\frac{\mathrm{d}u^k}{\mathrm{d}g^{\bar{\ell}}} = -1 + \Big[1 + \kappa^{\bar{\ell}}(n-1)\Big]\, V_G(G^*) = 0, \tag{S4}$$

and, due to $\kappa^{\bar{\ell}} < \kappa^k = \max\{\kappa^1, \ldots, \kappa^m\}$, for group k we obtain

$$\frac{\mathrm{d}u^k}{\mathrm{d}g^k} = -1 + \left[1 + \kappa^k(n-1)\right] V_G(G^*) > 0. \tag{S5}$$

Thus, the economy have not yet attained an equilibrium, because group k individuals have an incentive to further increase their contribution, a contradiction to the equilibrium assumption. In equilibrium we therefore have $G^* = n^k g^{k*}$ determined by

$$\frac{\mathrm{d}u^k}{\mathrm{d}g^k} = -1 + \left[1 + \kappa^k(n-1)\right] V_G(G^*) = 0. \tag{S6}$$

An increase in morality of all groups $\ell \neq k$ does not change the equilibrium public good provision G^* and increasing morality of group k increases public good provision G^*, since

$$\frac{\mathrm{d}G^*}{\mathrm{d}\kappa^k} = -\frac{(n-1)V_G(G^*)}{\left[1 + (n-1)\kappa^k\right] V_{GG}(G^*)} > 0, \tag{S7}$$

where we have applied the Implicit Function Theorem to (S6). In sum, with heterogeneity in altruistic preferences we obtain neither a polarization incentive of the individuals nor a detrimental effect of an increase in altruism on aggregate public good provision, in contrast to our analysis with heterogenous *homo moralis* preferences.

We can generalize this result to interior solutions with general utility functions $W(x)$ and $V(G)$. The first-order condition of the utility maximization of a group k individual then follows from setting (S2) equal to zero. We obtain

$$W_x\big(\omega - g^{k*}\big) = \left[1 + \kappa^k(n-1)\right] V_G(G^*). \tag{S8}$$

Using the inverse function of W_x, solving for g^{k*}, multiplying with n^k and finally summing over all groups gives an implicit function determining G^*, i.e.

$$G^* = \omega - \sum_{\ell=1}^{m} n^\ell W_x^{-1}\Big(\left[1 + \kappa^k(n-1)\right] V_G(G^*)\Big). \tag{S9}$$

Using again the Implicit Function Theorem as well as $(W_x^{-1})_x = 1/W_{xx}$, (S9) implies

$$\frac{\mathrm{d}G^*}{\mathrm{d}\kappa^k} = -\frac{\dfrac{n^k(n-1)V_G}{W_{xx}^k}}{1 + \displaystyle\sum_{\ell=1}^{m} \frac{n^\ell[1 + \kappa^\ell(n-1)]V_{GG}}{W_{xx}^\ell}} > 0, \tag{S10}$$

where $W_{xx}^\ell := W_{xx}\big(\left[1 + \kappa^\ell(n-1)\right] V_G(G^*)\big)$. Thus, an increase in altruism of group k again does not induce a polarization effect and always increases aggregate public good provision, in contrast to the model with heterogenous *homo moralis* individuals.

Fairness. With fairness preferences, the utility of an individual from group k reads

$$\begin{aligned} u^k &= W\big(\omega - g^k\big) + V\Bigg[g^k + (n^k-1)\bar{g}^k + \sum_{\ell=1,\ell\neq k}^{m} n^\ell g^\ell\Bigg] \\ &\quad - \kappa^k \Bigg\{ W\big(\omega - g^k\big) + V\Bigg[g^k + (n^k-1)\bar{g}^k + \sum_{\ell=1,\ell\neq k}^{m} n^\ell g^\ell\Bigg] \\ &\qquad - \frac{1}{n}\sum_{\ell=1}^{m} n^\ell \Bigg[W\big(\omega - g^\ell\big) + V\Bigg[g^k + (n^k-1)\bar{g}^k + \sum_{j=1,j\neq k}^{m} n^j g^j\Bigg]\Bigg]\Bigg\}^2. \\ &= W\big(\omega - g^k\big) + V\Bigg[g^k + (n^k-1)\bar{g}^k + \sum_{\ell=1,\ell\neq k}^{m} n^\ell g^\ell\Bigg] \\ &\quad - \kappa^k \Bigg\{ W\big(\omega - g^k\big) - \frac{1}{n}\sum_{\ell=1}^{m} n^\ell W\big(\omega - g^\ell\big)\Bigg\}^2, \end{aligned} \tag{S11}$$

where $\kappa^k > 0$ is now the degree to which the individual takes into account the quadratic fairness costs that are caused by the deviation of this individual's utility from average utility of all individuals. For simplicity, we assume that fairness costs are symmetric, i.e. a higher than average utility cause the same costs as a lower than average utility of the same size. Taking the derivatives of (S11) with respect to g^k yields

$$\begin{aligned} \frac{\mathrm{d}u^k}{\mathrm{d}g^k} &= -W_x\big(\omega - g^k\big) + V_G(G) \\ &\quad + 2\kappa^k \Bigg\{ W\big(\omega - g^k\big) - \frac{1}{n}\sum_{\ell=1}^{m} n^\ell W\big(\omega - g^\ell\big)\Bigg\} \frac{n-n^k}{n} W_x\big(\omega - g^k\big), \end{aligned} \tag{S12}$$

$$\frac{\mathrm{d}^2u^k}{\big(\mathrm{d}g^k\big)^2} = W^k_{xx} + V_{GG} + 2\kappa^k \Bigg\{ \Bigg[\frac{n-n^k}{n} W^k_x\Bigg]^2 - \Bigg[W^k - \frac{1}{n}\sum_{\ell=1}^{m} n^\ell W^\ell\Bigg] \frac{n-n^k}{n} W^k_{xx}\Bigg\}, \tag{S13}$$

where, for notational convenience, we use the short cuts $W^k := W\big(\omega - g^k\big)$, $W^k_x := W_x\big(\omega - g^k\big)$ and $W^k_{xx} := W_{xx}\big(\omega - g^k\big)$ in equation (S13).

In order to ensure comparable conditions to Proposition 3, we assume the utility function to be quasi-linear in the private good, i.e. $W_x = 1$ and $W_{xx} = 0$, as well as the quadratic sub-utility function (19). Furthermore, we focus on an interior solution to the individual's utility maximization problem. From (S12) and (S13), the corresponding first- and second-

order conditions can then be written as

$$\frac{\mathrm{d}u^k}{\mathrm{d}g^k} = -1 + \gamma_1 - \gamma_2 G - 2\kappa^k \left\{ g^k - \frac{1}{n}\sum_{\ell=1}^{m} n^\ell g^\ell \right\} \frac{n - n^k}{n} = 0, \tag{S14}$$

$$\frac{\mathrm{d}^2 u^k}{\left(\mathrm{d}g^k\right)^2} = -\gamma_2 + 2\kappa^k \left(\frac{n - n^k}{n} \right)^2 < 0. \tag{S15}$$

From (S15) we see that γ_2 has to be sufficiently large for the second-order condition to be satisfied. Using $G = \sum_{\ell=1}^{m} n^\ell g^\ell$ in (S14) and rearranging yields

$$g^k - \frac{G}{n} = \frac{n}{2\kappa^k (n - n^k)} \Big[\gamma_1 - 1 - \gamma_2 G \Big], \tag{S16}$$

Multiplying both sides with n^k and summing over all groups, the LHS of (S16) becomes zero and we obtain the equilibrium public good provision

$$G^* = \frac{\gamma_1 - 1}{\gamma_2}, \tag{S17}$$

where we implicitly assume $\gamma_1 > 1$ in order to ensure an interior solution. According to (S17), aggregate public good provision does not depend on fairness preferences κ^k at all. Hence, we obtain neither a polarization effect nor a detrimental effect of increasing fairness, in contrast to the framework with heterogenous *homo moralis* individuals.

Social norm. With a social norm regarding the contribution to the public good, the utility of an individual from group k can be written as

$$u^k = W\big(\omega - g^k\big) + V\left[g^k + (n^k - 1)\bar{g}^k + \sum_{\ell=1, \ell \neq k}^{m} n^\ell g^\ell \right] - \kappa^k \left\{ g^k - \frac{1}{n}\sum_{\ell=1}^{m} n^\ell g^\ell \right\}^2, \tag{S18}$$

where $\kappa^k > 0$ now measures the degree to which the individual takes into account the quadratic costs from the deviation of its contribution level to the social norm, which equals the average of all contribution levels. The derivatives of (S18) are

$$\frac{\mathrm{d}u^k}{\mathrm{d}g^k} = -W_x\big(\omega - g^k\big) + V_G(G) - 2\kappa^k \left\{ g^k - \frac{1}{n}\sum_{\ell=1}^{m} n^\ell g^\ell \right\} \frac{n - n^k}{n}, \tag{S19}$$

$$\frac{\mathrm{d}^2 u^k}{\left(\mathrm{d}g^k\right)^2} = W_{xx}\big(\omega - g^k\big) + V_{GG}(G) - 2\kappa^k \left(\frac{n - n^k}{n} \right)^2 < 0. \tag{S20}$$

According to (S20), the second-order condition of utility maximization is always satisfied. Setting (S19) equal to zero and using the quasi-linear specification with $W_x = 1$ and $V(G)$

given by (19) as in Proposition 3, the first-order condition to utility maximization is again represented by (S14). By the same steps as in case of fairness preferences, we therefore obtain the equilibrium public good provision in (S17), so there is again neither a polarization incentive nor a detrimental effect of increasing social preferences by increasing κ^k, in contrast to the heterogenous *homo moralis* framework.