Supplementary Document

Sufficientarianism and Egalitarianism

The ethical rule in our general Social Welfare Function (SWF) in Equation S15 can incorporate various principles of distributive justice beyond Utilitarian and Prioritarian views. Here, we use Sufficientarian and Egalitarian SWFs to calculate welfare. Sufficientarianism extends the Utilitarian SWF by introducing a sufficiency threshold corresponding to the international poverty line of $1.25 United States Dollars (adjusted to the 2005 Purchasing Power Parity) stipulated by the World Bank. The Egalitarian approach partially utilizes the Prioritarian formulation to satisfy the Fundamental Axioms of SWF (see Social Welfare Function Section). Additionally, the Egalitarian approach integrates a measure of relative inequality using the GINI index.

We select Pareto optimal policies by filtering the top 10% in the welfare objective and then choosing the best-performing policy on the climate objective. The Sufficientarian framework allows for growth in developing nations and does not advocate rapid near-term mitigation, unlike the Utilitarian approach (see Extended Figure 4). However, the Sufficientarian policy intensifies mitigation around 2050, surpassing Utilitarian levels while maintaining a similar regional distribution of the mitigation burden (Extended Figure 5). The Egalitarian policy suggests a mitigation pathway similar to the Prioritarian approach, stipulating significant reductions leading up to 2050. Nevertheless, due to its emphasis on the relative positions of regions (i.e. how equally welfare and welfare losses are distributed), the mitigation burden is more evenly spread (Extended Figure 5). This approach has a levelling-down effect, meaning it makes everyone worse off in an attempt to equalize welfare and welfare losses, which goes against the Common but Differentiated Responsibilities and Respective Capabilities principle adopted by the United Nations Framework Convention on Climate Change (UNFCCC).

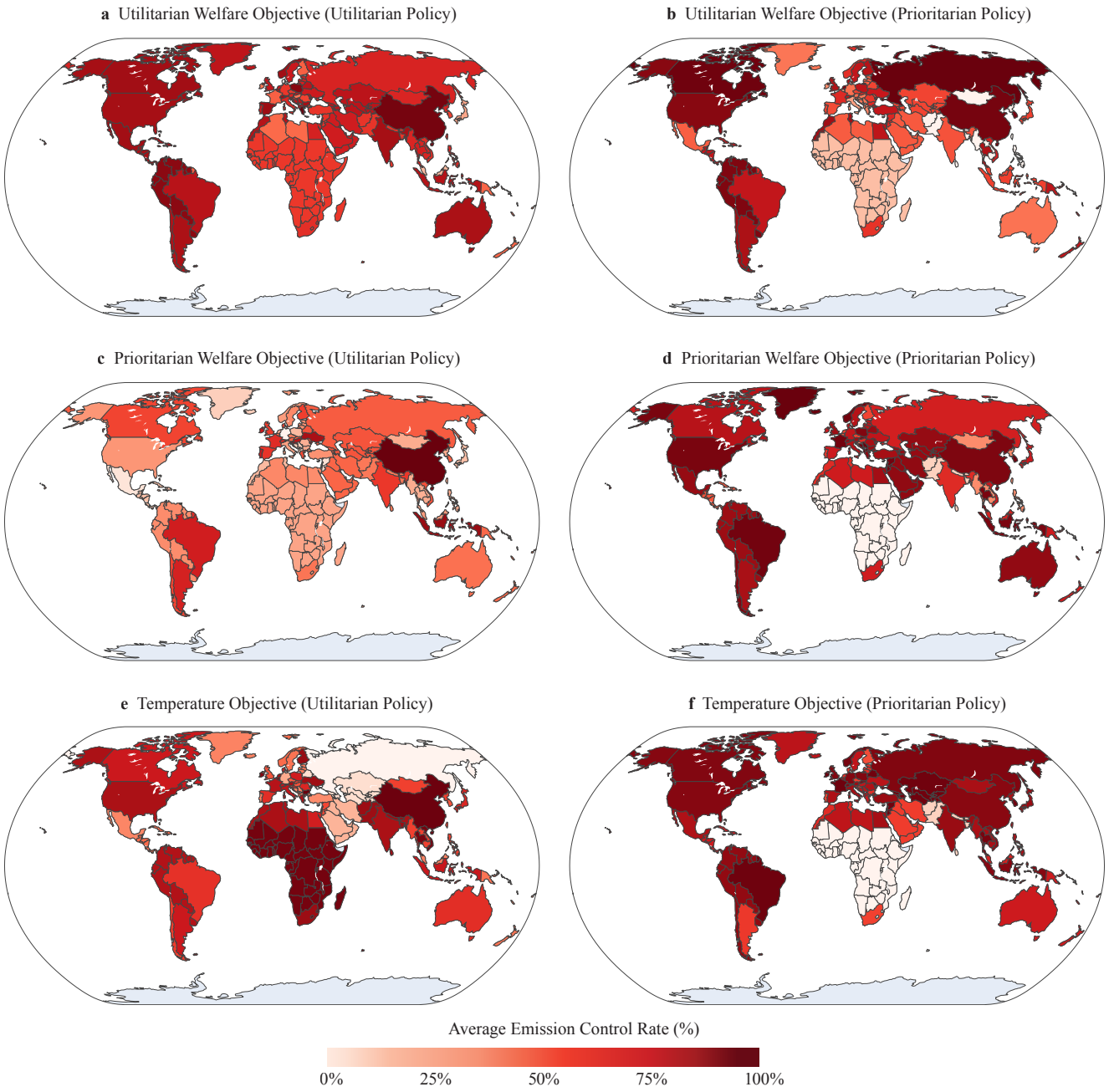
The Sufficientarian mitigation pathway results in the highest temperature rise by the end of the century because it does not support rapid near-term mitigation, as evidenced by the emission pathways shown in Extended Figure 4. Like the Prioritarian policy, the Egalitarian policy keeps warming below 2°C by recommending near-term rapid cuts in emissions. This rapid mitigation underscores the urgent need for near-term rapid and sustained mitigation actions on a global scale, as highlighted by IPCC reports, to maintain warming below 2°C. The comparison of emission and temperature pathways in Extended Figure 4 shows that ethical preferences and the selection of normative parameters substantially affect the optimal policy pathways.

Equity Implication of different solutions from the Pareto Set

Comparing Utilitarian and Prioritarian approaches requires reevaluation; a Prioritarian Pareto optimal policy must be assessed using a Utilitarian SWF and vice versa. Reevaluating Prioritarian policies with a Utilitarian SWF yields their Utilitarian welfare scores while reevaluating Utilitarian policies with a Prioritarian SWF produces their Prioritarian welfare scores. This reevaluation, depicted in Extended Data Figure 1 and 2, highlights a trade-off: increasing Prioritarian welfare reduces Utilitarian welfare. Additionally, higher Prioritarian welfare is positively correlated with fewer years above the temperature threshold, whereas Utilitarian welfare shows a negative correlation. We select the reevaluated Utilitarian and Prioritarian policies that achieve similarly high scores on three objectives—the Utilitarian welfare, the Prioritarian welfare, and the temperature objective—with minimal differences between their scores. We then assess the equity implications of these Pareto optimal solutions by separately comparing how each ethical framing distributes the mitigation burden for policies that perform similarly on each objective (marked with red stars in Extended Figure 2).

Supplementary Figure 1 compares the global distribution of the average emission control rate in 2050 under the Utilitarian and Prioritarian policies. In Extended Figure 2, each red star indicates the location of the two selected policies (one Utilitarian and one Prioritarian) chosen for intercomparison on that objective. All three stars appear near the top of their respective objective axes, which indicates that the selected policies score high on each objective. Each star represents a pair of solutions with minimal difference between their scores on that particular objective.

As seen in Supplementary Figure 1, the distribution pattern is distinct and persistent between Utilitarian and Prioritarian; only the intensity of the emission control rate differs based on the selected objective. The Utilitarian policy seeks to distribute the emission control rate evenly between developed and developing countries. In contrast, the Prioritarian policy consistently shifts the burden from developing and least-developed nations to developed ones. This inequity becomes more flagrant for the best-performing policies under the temperature objective. The Utilitarian policy places most of the burden on least-developed nations in Sub-Saharan Africa and South Asia while imposing little to no mitigation on Russia. In contrast, the Prioritarian

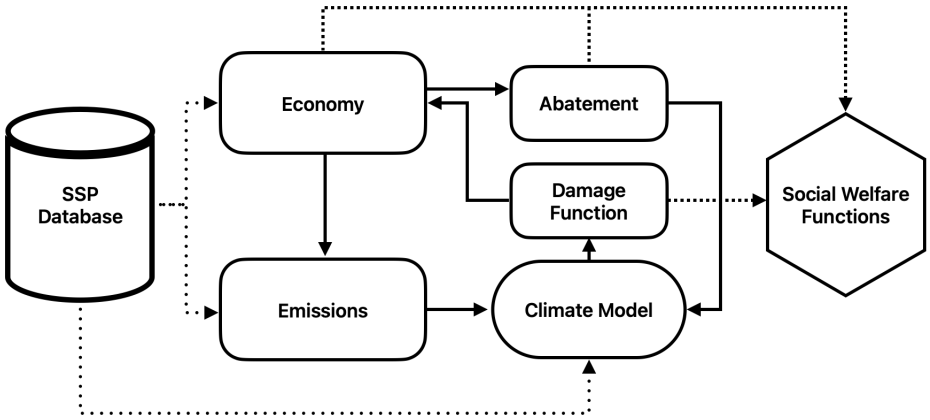


Supplementary Figure 1. Comparison of Global Distribution of Emission Control Rate in 2050 for Utilitarian and Prioritarian Policies Across Three Objectives. (**a**) Best Utilitarian Policy evaluated under the Utilitarian SWF. (**b**) Best Prioritarian Policy evaluated under the Utilitarian SWF. (**c**) Best Utilitarian Policy evaluated under the Prioritarian SWF. (**d**) Best Prioritarian Policy evaluated under the Prioritarian SWF. (**e**) Best Utilitarian Policy under the Temperature Objective. (**f**) Best Prioritarian Policy under the Temperature Objective.

solution reverses this pattern by shifting the burden to developed nations such as the United States, Canada, Russia, and Europe. Comparing these Pareto-optimal solutions across different objectives highlights the importance of disaggregation; prioritizing a single objective can obscure important co-benefits and trade-offs, thereby perpetuating existing injustices. This comparison underscores the need for a multiobjective approach to uncover hidden biases.

JUSTICE Integrated Assessment Model

The JUSTICE integrated assessment model (IAM) is a modular and efficient framework designed to evaluate various modelling assumptions related to economic growth, damage functions, abatement costs, and social welfare. It is a surrogate for more complex process-based IAMs, allowing testing under different uncertainties without prohibitively high computational costs. JUSTICE inherits the economy, damage, and abatement modules from the RICE50+ model[1]. These modules are integrated with the FaIR 2.1.0 climate model, enabling probabilistic assessments of climate policies under different climate sensitivity scenarios. In contrast to other IAMs, JUSTICE searches for Pareto-optimal policies within a multi-objective framework while also evaluating the robustness of these policies regarding their environmental, economic, and distributional impacts across regions under diverse socioeconomic and climate scenarios. Socioeconomic drivers, including population growth and economic development at the regional level, are drawn from the five Shared Socioeconomic Pathways (SSP), enabling the model to investigate five assumptions across alternative futures for socioeconomic development. By balancing trade-offs between multiple objectives and considering different ethical perspectives, the model facilitates the exploration of normative uncertainty and aims to formulate equitable climate policies.



Supplementary Figure 2. Overview of JUSTICE IAM

FaIR Climate Model

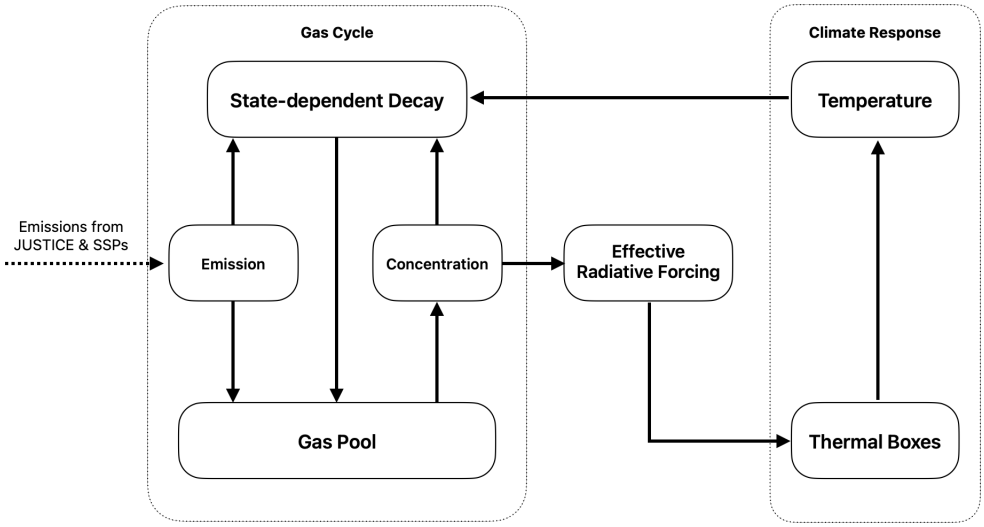
The Finite-amplitude Impulse Response (FaIR) is a simplified climate model designed to emulate the behaviour of more complex Earth System Models from the Coupled Model Intercomparison Project Phase 6 (CMIP6). FaIR employs a computationally efficient six-equation model to simulate the global mean climate response to 52 greenhouse gases (GHGs) from 1750 to 2100[2]. The GHG emissions used to calculate the concentration come from the total CO2 emissions, calculated by summing the emissions from fossil fuel and industrial processes modelled by JUSTICE IAM and the CO2 emissions from agriculture, forestry, and other land use (AFOLU). The AFOLU emissions, along with other GHG emissions, volcanic activity, and solar forcing data, are exogenous in this study; these data come from the SSP and RCP dataset obtained via the Reduced Complexity Model Intercomparison Project (RCMIP)[3].

FaIR consists of six equations: five that form the standard impulse model used for GHG metrics in the Intergovernmental Panel on Climate Change's (IPCC) Fifth Assessment Report and one additional equation to represent state-dependent feedback, which accounts for nonlinearities in the carbon cycle. These nonlinearities stem from atmospheric decay feedback within carbon and methane cycles. The model operates by taking GHG emissions to derive concentrations, which are then used to calculate the effective radiative forcing. Finally, the climate system's temperature response is computed from this forcing. The processes involved can be broadly categorized into three components: gas cycle, radiative forcing, and climate response.

FaIR's computational efficiency facilitates the exploration of climate uncertainty via a constrained calibrated ensemble, where climate uncertainty is quantified using Equilibrium Climate Sensitivity (ECS) and Transient Climate Response (TCR). The constrained posterior ensemble was selected from a prior ensemble of 1.5 million members, which sampled uncertainties in the carbon cycle, effective radiative forcing, and thermal response. This full ensemble was consistent with IPCC's estimated ranges of ECS and TCR and historical climate data, including ocean heat uptake, global mean surface temperature, aerosol effective radiative forcing, carbon dioxide concentrations, and projected future warming from the SSP2-RCP4.5 scenario. The

constrained posterior sample of 1,001 members was selected based on the alignment of each member's probability with the likelihood of present-day warming. The constrained ensemble members successfully reproduce historically observed warming, with each member representing a potential "state of the world" (denoted by the set S) where the social planner is assumed to have perfect knowledge of the climate system's state[4]. FaIR calculates the global mean temperature increase, $\mathbf{GMT}_{t_i S}$, for time step t_i across all states in S by summing the regional emissions from the previous time step t_{i-1}, as shown in Equation S1.

$$\mathbf{GMT}_{t_i S} = FAIR\left(\sum_{n \in N} \mathbf{EMIS}_{t_{i-1} n_j S}\right), \quad \forall t_i \in T,\ \forall n_j \in N \tag{S1}$$



Supplementary Figure 3. Overview of FaIR

Social Welfare Function

Policies affect multiple dimensions of well-being, such as health, income, and the environment, and their effects on individuals can vary widely. Thus, a challenging aspect of policymaking involves balancing trade-offs between reducing inequality and enhancing overall welfare. The Social Welfare Framework (SWF) provides a robust mechanism to ethically evaluate government policies by considering their multidimensional and heterogeneous impacts and balancing conflicting trade-offs. Primarily a normative framework, the SWF assesses and ranks policy options based on ethical considerations. The SWF adopts welfare-consequentialism to evaluate policies. It is welfarist because it determines the ranking of outcomes based on the welfare patterns of individuals in different scenarios. It is also consequentialist because it evaluates policies based on the consequences of these outcomes.

The SWF comprises a two-step process to calculate the ranking of policy outcomes. The first step is to convert bundles of attributes (both income and non-income attributes that are used as a measure of well-being) into a welfare measure. A welfare measure is a mathematical indicator of an individual's welfare, calculated using a utility function—a mathematical tool that embodies the decision-maker's preferences (see Equation S2). It is a prerequisite that this welfare measure accommodates both inter- and intrapersonal welfare comparisons, including welfare levels and differences. The second step is to apply an ethical rule based on distributive justice principles to rank the welfare distribution among individuals. The rule for ranking welfare patterns must satisfy the fundamental ethical axioms of the SWF framework. These widely accepted axioms clarify the ethical merits of different SWFs, each rooted in distinct principles of distributive justice. The fundamental axioms include:

1. **Pareto Principles:** The Pareto principles comprise two sub-principles—Pareto Indifference and Strong Pareto.

   (a) **Pareto Indifference:** Two outcomes are ethically equal if no one's welfare changes between them.

(b) **Strong Pareto:** An outcome is ethically better if it improves at least one individual's welfare without harming anyone else's.

2. **Anonymity:** This principle embodies impartiality, stipulating that everyone's welfare is equal, regardless of who they are or where and when they live.

3. **Principle of Invariance:** The ranking of outcomes should remain unchanged when the welfare measure is rescaled (for example, using a mathematical transformation), as long as both the levels and differences in welfare are preserved.

In addition to the fundamental axioms, the additional axioms reflect different ethical perspectives and are more contested. Different SWFs may adhere to different sets of these additional axioms, as their acceptance depends on individual ethical viewpoints—none of which can be deemed definitively correct. These include:

4. **Pigou-Dalton:** Emphasizes fairness in welfare distribution. It considers any transfer that reduces the gap between a better-off and a worse-off person as an ethical improvement provided that total welfare remains constant.

5. **Separability:** Simplifies ethical evaluation by focusing solely on individuals affected by the change, ignoring those unaffected.

6. **Continuity:** Requires a continuous function representing the SWF's ethical rule that converts welfare patterns into a real number. Continuity ensures the selection of the best outcome, even when the outcome set is infinite—crucial for decision-making under uncertainty.

The SWF approach can be extended to different dimensions when addressing complex problems such as climate change. Climate change policymaking involves addressing intragenerational justice (how welfare is distributed between regions), intertemporal concerns (how welfare is distributed among generations), and designing robust policies under deep uncertainty. Thus, inequities can arise in the spatial, temporal, and uncertainty dimensions. Although these dimensions of inequities can be intertwined, Atkinson et al.[5] empirically found that inequity preferences are distinct and vary substantially across these three dimensions. Based on their survey, they discovered that inequity aversion is highest in the spatial dimension, followed by the temporal dimension, and lowest in the uncertainty dimension.

Berger and Emmerling[6] extended the SWF to apply across these three dimensions by calculating welfare based on equity equivalents. The von Neumann-Morgenstern (vNM) utility function represents equity preferences in a particular dimension. The vNM utility function is described in Equation S2 where D is the set of dimensions and M is a generic element from D. The set T is an ordered set containing discrete time steps, set N representing the individual regions, and set S contains the ensemble members representing the states of the world. The $\mathbf{C}_M$ is the unit distributed along dimension M, and φ_M is the inequity aversion parameter for dimension M.

$$f(\mathbf{C}_M;\varphi_M) = \begin{cases} \frac{\mathbf{C}_M^{1-\varphi_M}}{1-\varphi_M}, & \text{if } \varphi_M \neq 1 \\ \ln(\mathbf{C}_M), & \text{if } \varphi_M = 1 \end{cases} \quad \text{where } M \in D = \{S,N,T\} \tag{S2}$$

The inequity aversion parameter φ_M captures the inequity aversion of the social planner on dimension M. The social welfare is equal to the equity equivalent $\mathscr{E}^M$ in dimension M and is calculated using an attribute bundle (a metric that acts as a proxy of well-being), or simply consumption per capita (CPC), $\mathbf{C}_M$ by applying the Equation S3. The $\mathbb{E}^M$ is an aggregator of utilities in dimension M. Aggregation over a certain dimension collapses that dimension, and inverting this aggregated sum of utilities with the inverse function f returns equity equivalent CPC, $\bar{\mathbf{C}}$. Equity equivalent in dimension M signifies the level of CPC that, when uniformly distributed across dimension M, does not change the utility score (See Berger & Emmerling[6] for a detailed explanation of equity equivalents). Following the conventions established in previous IAM studies, we use CPC, $\mathbf{C}_{TNS}$, as a proxy for welfare. The welfare score, w, is computed from the $\mathbf{C}_{TNS}$ hypermatrix using Equation S5. This welfare score is a scalar and is derived after applying the aggregator function over the three dimensions, which collapses the hypermatrix into a scalar welfare score.

$$\widetilde{\mathbf{C}} = \mathscr{E}^M(\mathbf{C}_M) = f^{-1}\mathbb{E}^M[f(\mathbf{C}_M;\varphi_M)] \tag{S3}$$

$$\mathbb{E}^M[f(\mathbf{C}_M;\varphi_M)] = \sum_{M\in D} q_M[f(\mathbf{C}_M;\varphi_M)] \tag{S4}$$

In Equation S4, the weight q_M has different interpretations based on the dimension under consideration. Equations S2, S3 and S4 are general equations applied across the three dimensions, but the interpretations of the utility function, inequity aversion parameter, the weight and the equity equivalent change based on the dimension being considered.

$$w = \mathscr{E}^T\left(\mathscr{E}^N\left(\mathscr{E}^S(\mathbf{C}_{TNS})\right)\right) \tag{S5}$$

When considering all three dimensions, the order of aggregation over the dimensions matters if preferences over them are different, as mathematically proved by Berger and Emmerling[6]. In our case, where there are 57 different regions (N), 286 years (T), and 1001 states (S) from FaIR climate ensemble members, the CPC $\mathbf{C}_{TNS}$ is a 3D hypermatrix. The aggregation starts with the uncertainty dimension, followed by the spatial dimension, and it ends with the temporal dimension, where time discounting is applied. This order of aggregation follows an individual risk aversion approach, where the social planner considers the risk of CPC loss for individual regions in different states-of-the-worlds (SOWs) and aggregates all the region's certainty equivalent CPC, represented as $\widetilde{\mathbf{C}}_{TN}$ (See Equation S7). The set (T) is an ordered set containing discrete time steps (Equation S6), set (N) representing the individual regions (Equation S6a), and set (S) contains the ensemble members representing the states of the world (Equation S6b).

$$T = \{t_0, t_1, \ldots, t_i\} \text{ where } t_0 < t_1 < \ldots < t_i \tag{S6}$$

$$N = \{n_1, n_2, \ldots, n_j\} \tag{S6a}$$

$$S = \{s_1, s_2, \ldots, s_k\} \tag{S6b}$$

Uncertainty dimension $(M = S)$ addresses the risk of CPC loss across different SOWs, with the risk attitude expressed by the risk aversion parameter φ_S. The vNM utility function captures the social planner's risk attitude using this parameter (Equation S7b). The weight q_{s_k} is the probability associated with the state s_k. The equity equivalent $\mathscr{E}^S$ represents the certainty equivalent of CPC, which is the CPC value if the CPC is uniformly distributed across the SOWs. If $\varphi_S > 0$, extra priority is given to the region with the lowest CPC level in any SOW. For this study, we assume $\varphi_S = 0$, implying a risk-neutral decision maker. In this case, the equity equivalent CPC, $\widetilde{\mathbf{C}}_{TN}$, becomes the average across the different SOWs. Since CPC over different SOWs is aggregated, $\widetilde{\mathbf{C}}_{TN}$ is a 2D matrix containing only the T and N dimensions. Being risk-neutral offers simplicity in terms of satisfying the SWF axioms. If the decision maker is risk-averse or risk-seeking, additional uncertainty-related axioms will apply, depending on the ethical premise[7].

$$\widetilde{\mathbf{C}}_{TN} = \mathscr{E}^S(\mathbf{C}_{TNS}) = f^{-1}\left[\left(\mathbb{E}^S\left[f\left(\mathbf{C}_{TNs_k};\varphi_S\right)\right]\right);\varphi_S\right] \tag{S7}$$

$$\mathbb{E}^S\left[f\left(\mathbf{C}_{TNs_k}\right)\right] = \sum_{s_k\in S} q_{s_k} f\left(\mathbf{C}_{TNs_k}\right) \tag{S7a}$$

$$f\left(\mathbf{C}_{TNs_k};\varphi_S\right) = \begin{cases} \frac{{\mathbf{C}_{TNs_k}}^{1-\varphi_S}}{1-\varphi_S} & \text{if } \varphi_S \neq 1 \\ \ln\left(\mathbf{C}_{TNs_k}\right) & \text{if } \varphi_S = 1 \end{cases} \quad \forall s_k \in S \tag{S7b}$$

Region Dimension $(M = N)$ is where different principles of distributive justice apply. The utility score in this dimension is based on the "ethical rule" and how this rule values the distribution pattern of welfare across different regions n_j at any given time step t_i. The preference parameter for the vNM utility function is the inequality aversion parameter φ_N, also known as Atkinson's priority parameter, which prioritizes the CPC of the worse-off regions when $\varphi_N > 0$. The weights q_{n_j} in Equation S8a represent the ratio of the region n_j's population to the total world's population. The equity equivalent in the regional

dimension, $\mathscr{E}^N$, is equivalent to the region aggregated consumption $\widetilde{\mathbf{C}}_T$ and corresponds to an equally distributed equivalent. Equally distributed equivalent means that if CPC were equally distributed across all regions, the aggregated utility score would be the same as that CPC value. An additional transformation that applies in the region dimension is the concept of declining marginal utility of consumption. This transformation is also carried out using the vNM utility function, where the preference parameter becomes the elasticity of marginal utility of consumption λ, as shown in Equation S8c. Declining marginal utility ensures that an increase in CPC by a specific amount has a smaller impact if the region is already at a high level of CPC compared to a region at a lower level. Aggregating the regional CPC over the region dimension N collapses the dimension, resulting in a one-dimensional array—the global utility scores over all the time steps, $\mathbf{U}_T$. This utility score depends on the chosen distributive justice principle that informs the ethical rule.

$$\widetilde{\mathbf{C}}_T = \mathscr{E}^N\left(\widetilde{\mathbf{C}}_{TN}\right) = f^{-1}\left[\left(\mathbb{E}^N\left[f\left(\widetilde{\mathbf{C}}_{Tn_j};\varphi_N\right)\right]\right);\varphi_N\right] \tag{S8}$$

$$\mathbb{E}^N\left[f\left(\widetilde{\mathbf{C}}_{Tn_j}\right)\right] = \sum_{n_j \in N} q_{n_j} f\left(\widetilde{\mathbf{C}}_{Tn_j}\right) \tag{S8a}$$

$$f\left(\widetilde{\mathbf{C}}_{Tn_j};\varphi_N\right) = \begin{cases} \frac{\widetilde{\mathbf{C}}_{Tn_j}^{1-\varphi_N}}{1-\varphi_N} & \text{if } \varphi_N \neq 1 \\ \ln\left(\widetilde{\mathbf{C}}_{Tn_j}\right) & \text{if } \varphi_N = 1 \end{cases} \quad \forall n_j \in N \tag{S8b}$$

$$\mathbf{U}_T = g\left(\widetilde{\mathbf{C}}_T;\lambda\right) = \begin{cases} \frac{\widetilde{\mathbf{C}}_T^{1-\lambda}}{1-\lambda} & \text{if } \lambda \neq 1 \\ \ln\left(\widetilde{\mathbf{C}}_T\right) & \text{if } \lambda = 1 \end{cases} \tag{S8c}$$

Utilitarianism, which traces back to Jeremy Bentham, is the most widely used ethical rule in SWFs. In the Utilitarian SWF, the inequality aversion parameter is zero ($\varphi_N = 0$), making the rule a simple sum of the utilities across regions. Utilitarian SWF must be used with CPC, which has been transformed with declining marginal utility. Without this transformation, a Utilitarian SWF would be indifferent to whether an increase in consumption occurs in a worse-off or well-off region. A Utilitarian SWF with declining marginal utility satisfies all the fundamental and additional axioms mentioned earlier except for the Pigou-Dalton principle. Consequently, the Utilitarian SWF remains insensitive to the distribution of welfare. The Utilitarian SWF on the spatial dimension is shown in Equation S9, where g is the vNM utility function that transforms the $\widetilde{\mathbf{C}}_{Tn_j}$ with declining marginal utility λ.

$$\mathbf{U}_T^{\text{UTILITARIAN}} = g\left(\mathbb{E}^N\left[\widetilde{\mathbf{C}}_{Tn_j}\right];\lambda\right) \tag{S9}$$

Prioritarianism, or continuous Prioritarian SWF, applies a strictly increasing and concave transformation function to the 2D utility matrix $\widetilde{\mathbf{C}}_{TN}$ before aggregating regional welfare scores. This additional transformation makes the Prioritarian SWF sensitive to the distribution of welfare, unlike the Utilitarian SWF, and thereby satisfies the Pigou-Dalton principle. In this study, we use a transformation function from the Atkinson family because it satisfies the Principle of Invariance axiom. The Prioritarian SWF on the region dimension is shown in Equation S10), where g applies the declining marginal utility, and f applies the priority transformation function.

$$\mathbf{U}_T^{\text{PRIORITARIAN}} = g\left(f^{-1}\left[\mathbb{E}^N\left[f\left(\widetilde{\mathbf{C}}_{Tn_j};\varphi_N\right)\right];\varphi_N\right];\lambda\right) \tag{S10}$$

Sufficientarianism posits that it is ethically unacceptable for anyone's welfare to fall below a specified minimum threshold. This approach lies roughly between Utilitarianism and Prioritarianism. Below this welfare threshold, the Sufficientarian SWF behaves like a Prioritarian SWF; above the threshold, it behaves like a Utilitarian SWF. However, switching between Utilitarian and Prioritarian rules based on a threshold violates the continuity axiom and complicates inter-comparisons. To address this, we extend the Utilitarian SWF by incorporating a CPC threshold $\mathbf{C}_{TN}^{\text{Thresh}}$, as seen in Equation S11 and S12. This threshold acts as an asymptote during the welfare calculation for each region. The domain of Equation S11 changes from $[x \in \mathbb{R} \colon c_{t_i n_j} > 0]$ to $x \in \mathbb{R} \colon c_{t_i n_j} > c_{t_i n_j}^{\text{thresh}}$. If any region is below the threshold, the welfare value drops to negative infinity. Using this SWF ensures the optimizer keeps each region's welfare above this threshold. This modification satisfies the continuity principle, although

Sufficientarianism, like Utilitarianism, violates the Pigou-Dalton principle. However, it does satisfy a less robust version of Pigou-Dalton, described by Adler[7] as minimal Pigou-Dalton. For our Sufficientarian SWF, we chose the Sufficientarian threshold to be the World Bank's stipulated international poverty line of 1.25 USD per day ($c^{\text{thresh}} = 1.25$ measured in 2005 Purchasing Power Parity).

$$\mathbf{U}_T^{\text{SUFFICIENTARIAN}} = \; g\left(\mathbb{E}^N\left[\widetilde{\mathbf{C}}_{Tn_j} - \mathbf{C}_{Tn_j}^{Thresh}\right];\lambda\right) \tag{S11}$$

$$\mathbf{C}_{TN}^{\text{Thresh}} = \left\{c_{t_i n_j}^{\text{thresh}}\right\}, \quad \text{where} \quad c_{t_i n_j}^{\text{thresh}} = c^{\text{thresh}} \quad \forall t_i, n_j \tag{S12}$$

Egalitarianism focuses on the relative position of regional welfare, i.e., comparing the welfare of one region to all other regions in the world. Unlike Prioritarianism, which emphasizes improving the welfare of the worse-off, strict Egalitarianism cares about the equality of distribution, thus violating the separability axiom. Egalitarians intrinsically value the equality of welfare and rank outcomes based on how equally welfare is distributed. One major issue with this approach is the "levelling-down" objection, where a perfectly equal yet lower welfare distribution is ranked higher than an unequal but higher welfare distribution[8] Levelling-down violates the Pareto Principle, which states that any change that benefits at least one person without harming anyone else should be considered an improvement. To address this, Peterson and Hansson[9] introduced the Equality-Prioritarian SWF by extending the works of Temkin[10] and Rabinowicz[11]. Equality-Prioritarian SWF combines Prioritarianism with Egalitarianism by incorporating an inequality metric, such as the GINI index, and applying a Prioritarian transformation. This approach ensures that both the absolute welfare levels and the distribution of welfare matter. Equality-Prioritarianism or simply Egalitarian SWF used in our study is shown in Equation S13. It extends the Prioritarian SWF (Equation S10) by including an inequality metric, which modifies the welfare score based on distribution. The term $(1 - \text{GINI}_T(\widetilde{\mathbf{C}}_{TN}))$ weighs the welfare score such that higher inequality (GINI approaching 1) reduces the final welfare score and vice versa. We introduce an additional parameter called the egality strictness, ψ, to control the weighing of the final welfare score. The egality strictness parameter allows for adjusting or removing the concern for relative welfare among regions.

$$\mathbf{U}_T^{\text{EGALITARIAN}} = g\left(f^{-1}\left[\left(\mathbb{E}^N\left[f\left(\widetilde{\mathbf{C}}_{Tn_j};\varphi_N\right)\right]\cdot\left(1-\psi\cdot\text{GINI}_T\left(\widetilde{\mathbf{C}}_{TN}\right)\right)\right);\varphi_N\right];\lambda\right) \tag{S13}$$

We use the GINI metric to calculate the inequality of welfare distribution based on Concept-1 GINI, as described by Milanovic[8]. The Concept-1 GINI measures international inequality by comparing the distribution of CPC between countries or regions. In our study, regions are the units of observation, and CPC is used as a proxy for each region's welfare. The choice of metric for calculating inequality and the currency conversion method significantly impacts the calculated inequality. We use Purchasing Power Parity (PPP) in this study. We also assume constant consumption baskets across intra- and inter-generational comparisons to allow for consistent comparison. Concept-1 GINI indicates whether nations' CPC levels are converging. Equation S14 shows how GINI is computed by comparing differences in CPC levels between regions. Here, η represents the mean of the regional CPC (excluding regional population weights), $|N|$ is the total number of regions (57 in this study), and $\widetilde{\mathbf{C}}_{Tn_a}$ and $\widetilde{\mathbf{C}}_{Tn_b}$ are the CPC for regions a and b, respectively.

$$\text{GINI}_T\left(\widetilde{\mathbf{C}}_{TN}\right) = \text{GINI}\left(\widetilde{\mathbf{C}}_{Tn_j}\right) = \frac{1}{\eta}\cdot\frac{1}{|N|^2}\sum_{a=1}^{|N|}\sum_{b>i}^{|N|}\left|\widetilde{\mathbf{C}}_{Tn_b} - \widetilde{\mathbf{C}}_{Tn_a}\right| \tag{S14}$$

The General SWF of Distributive Justice Principles is derived by combining Equations S9, S10, S11, and S13 to accommodate all four principles on the Region dimension. The General SWF, as shown in equation S15, explicates the normative preferences of the different distributive justice principles into normative parameters that can be adjusted to fit the preferences of the decision-maker.

$$\mathbf{U}_T^{\text{GENERAL}} = g\left(f^{-1}\left[\left(\mathbb{E}^N\left[f\left(\widetilde{\mathbf{C}}_{Tn_j};\varphi_N\right)\right]\cdot\left(1-\psi\cdot\text{GINI}_T\left(\widetilde{\mathbf{C}}_{TN}\right)\right)\right);\varphi_N\right];\lambda\right), \quad \text{where} \;\; 0\le\psi\le 1 \tag{S15}$$

Time Dimension ($M = T$) is the last dimension to aggregate. The equity equivalent $\mathscr{E}^T(\mathbf{U}_T)$ in this dimension represents an equally distributed CPC level over time that yields the same intertemporal utility as the actual CPC stream (Equation S16). Intertemporal utilities are transformed using the function h in Equation S16b, with the intergenerational inequity aversion parameter, commonly known as the pure rate of time preference φ_T. The weight q_{t_i} in Equation S16a represents the discounted intertemporal population ratio of a region. Aggregation over long time horizons using the pure rate of time preference is known as time discounting, a method that has been heavily criticized and debated[12]. The debate includes opposing views from Nordhaus and Stern on the appropriate magnitude of discounting and arguments from scholars advocating for the total removal of time discounting[13]. Notably, any form of discounting violates the fundamental axiom of Anonymity. Discounting reduces the importance placed on future generations' welfare, making it ethically contentious within the framework of SWFs. Despite this contention, time discounting is widely used in the IAM literature, particularly with Utilitarian SWFs. For this reason, we retained discounting in both Utilitarian and Sufficientarian SWFs.

$$w = \mathscr{E}^T(\mathbf{U}_T) = \mathbb{E}^T\left[h(\mathbf{U}_{t_i};\varphi_T)\right] \tag{S16}$$

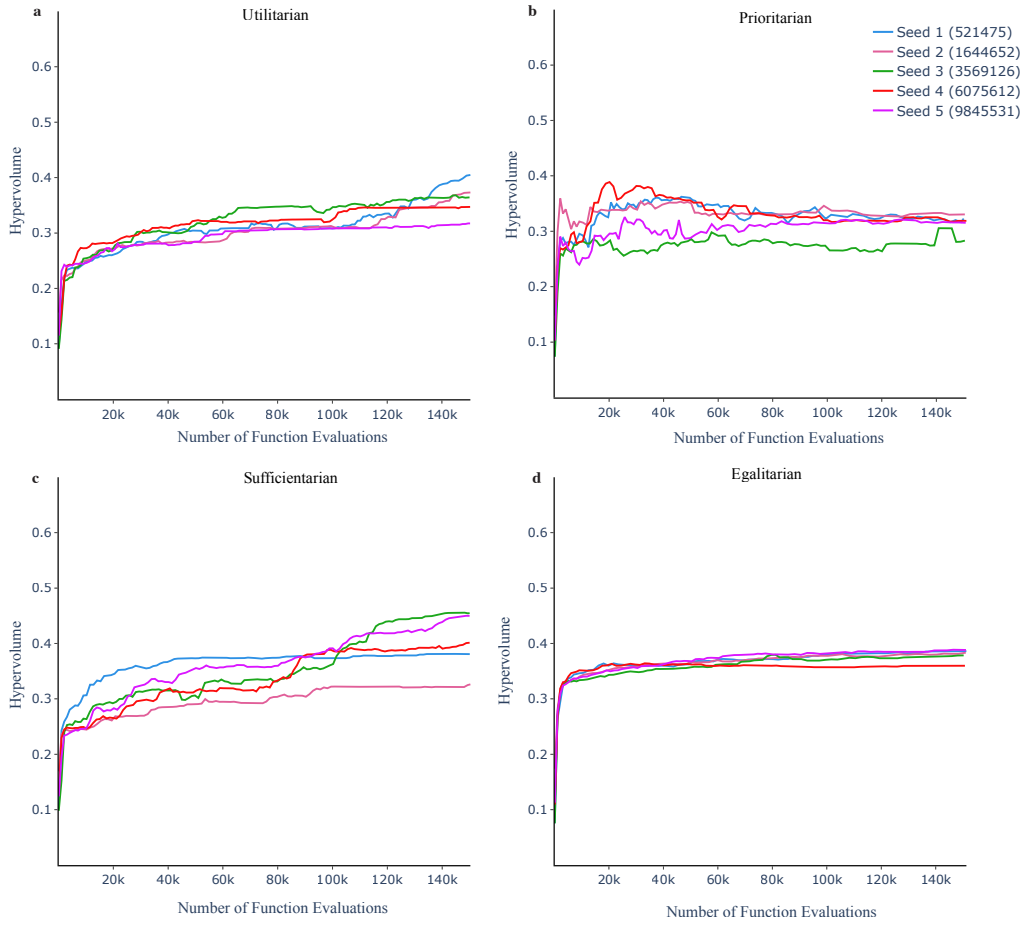
$$\mathbb{E}^T\left[h(\mathbf{U}_{t_i};\varphi_T)\right] = \sum_{t_i \in T} q_{t_i} h(\mathbf{U}_{t_i};\varphi_T) \tag{S16a}$$

$$h(\mathbf{U}_{t_i};\varphi_T) = \mathbf{U}_{t_i} \cdot (1+\varphi_T)^{-t} \quad , \quad \forall t_i \in T \tag{S16b}$$

Evolutionary Multiobjective Direct Policy Search

Evolutionary Multiobjective Direct Policy Search (EMODPS) combines Direct Policy Search (DPS) with Multi-Objective Evolutionary Algorithms (MOEAs) to optimize policy levers parameterized by non-linear activation networks, such as a Radial Basis Function (RBF)[14]. In this setup, the RBF maps system states (temperature feedback from JUSTICE) to control actions (emission control rate policy lever) within a closed-loop control problem. The MOEA tunes the RBF network hyperparameters and produces a Pareto front of Pareto-optimal policies, with each policy defined by a set of hyperparameters (centres, weights, and radii). This approach is particularly useful for high-dimensional, complex policy problems, such as optimizing climate policy for various regions over long time horizons and under climate uncertainty, as it directly optimizes policy hyperparameters while handling conflicting objectives and generating robust, diverse solutions.

We use the ε-NSGA-II Multiobjective Evolutionary Algorithm to optimize the hyperparameters of the RBF. This generational algorithm extends NSGA-II by incorporating ε-dominance, which divides the solution space into grid cells of size ε, allowing only one solution per cell to prevent similar solutions from dominating the Pareto front[15] while maintaining solution diversity and computational efficiency. Like NSGA-II, ε-NSGA-II also ranks candidate solutions based on Pareto dominance and crowding distance. We evaluate the quality and diversity of the Pareto fronts using hypervolume and epsilon values, choosing epsilon values for our objectives based on their numerical ranges (welfare = 0.1, years above temperature threshold = 0.25, welfare loss from damages = 10, welfare loss from abatement costs = 10). Given the algorithm's stochastic nature, which may cause sensitivity in the Pareto-optimal solutions to the random seed, we run multiple seeds (9845531, 1644652, 3569126, 6075612, 521475) to address this concern. We then test the growth of hypervolume over a series of function evaluations (150,000 evaluations for each seed and ethical framing) for all four problem formulations (see Supplementary Figure 4) and observe convergence in each case. Finally, we merge all candidate solutions from different runs and perform non-dominated sorting to select a reference set for our analysis.



Supplementary Figure 4. Hypervolume convergence over Number of Function Evaluations for four SWFs, each with five seed runs. (**a**) Utilitarian SWF. (**b**) Prioritarian SWF. (**c**) Sufficientarian SWF. (**d**) Egalitarian SWF.

Regional Configuration

Supplementary Table S1. JUSTICE Regional Configuration: List of Macro Regions, JUSTICE Regions, and Corresponding Countries

9 Macro Regions	JUSTICE Regions	Countries
United States	United States of America	United States
India	India	India
China	China	China
Other High Income	Australia	Australia
	Canada	Canada
	South Korea	South Korea
	Japan	Japan
Europe	Austria	Austria
	Belgium	Belgium
	Baltic States	Estonia, Latvia, Lithuania
	Denmark	Denmark
	Spain	Spain
	Finland	Finland
	France	France
	United Kingdom	United Kingdom
	Greece	Greece
	Hungary	Hungary
	Ireland	Ireland
	Italy	Italy
	Netherlands	Netherlands
	Norway	Norway
	Poland	Poland
	Portugal	Portugal
	Czechia	Czech Republic
	Germany	Germany
	Slovakia	Slovakia
	Slovenia	Slovenia
	Switzerland	Switzerland
	Sweden	Sweden
	Bulgaria	Bulgaria
	Croatia	Croatia

	Other European Countries	Åland Islands, Albania, Andorra, Bosnia and Herzegovina, Cyprus, Faroe Islands, Gibraltar, Greenland, Guernsey, Iceland, Isle of Man, Jersey, Kosovo, Liechtenstein, Luxembourg, Malta, Monaco, Montenegro, North Macedonia, San Marino, Serbia, Svalbard and Jan Mayen, Vatican City
	Romania	Romania
South and Southeast Asia	Indonesia	Indonesia
	Malaysia	Malaysia
	Other Southeast Asia	Brunei Darussalam, Cambodia, Cocos (Keeling) Islands, Hong Kong, Lao People's Democratic Republic, Macao, Mongolia, Myanmar, North Korea, Philippines, Singapore, Taiwan
	Rest South Asia	Afghanistan, Bangladesh, Bhutan, Maldives, Nepal, Pakistan, Sri Lanka
	Thailand	Thailand
	Vietnam	Vietnam
	Rest Pacific Islands	American Samoa, Christmas Island, Cook Islands, Fiji, French Polynesia, Guam, Heard Island and McDonald Islands, Kiribati, Marshall Islands, Micronesia, Nauru, New Caledonia, New Zealand, Niue, Norfolk Island, Northern Mariana Islands, Palau, Papua New Guinea, Pitcairn, Samoa, Solomon Islands, Tokelau, Tonga, Tuvalu, United States Minor Outlying Islands, Vanuatu, Wallis and Futuna
Gulf Countries	Gulf Countries	Bahrain, Iran, Iraq, Kuwait, Oman, Qatar, Saudi Arabia, United Arab Emirates, Yemen
Sub-Saharan Africa	Sub-Saharan Africa	Angola, Benin, Botswana, British Indian Ocean Territory, Bouvet Island, Burkina Faso, Burundi, Cameroon, Cape Verde, Central African Republic, Chad, Comoros, Congo, Democratic Republic of the Congo, Côte d'Ivoire, Djibouti, Equatorial Guinea, Eritrea, Eswatini, Ethiopia, French Southern and Antarctic Lands, Gabon, Gambia, Ghana, Guinea, Guinea-Bissau, Kenya, Lesotho, Liberia, Madagascar, Malawi, Mali, Mauritania, Mauritius, Mayotte, Mozambique, Namibia, Niger, Nigeria, Réunion, Rwanda, Saint Helena, São Tomé and Principe, Senegal, Seychelles, Sierra Leone, Somalia, South Sudan, Sudan, Tanzania, Togo, Uganda, Zambia, Zimbabwe
Rest of the World	Argentina	Argentina
	Brazil	Brazil
	Chile	Chile
	Mexico	Mexico

Rest Central America	Anguilla, Antigua and Barbuda, Aruba, Bahamas, Barbados, Belize, Bermuda, Bonaire, Sint Eustatius and Saba, British Virgin Islands, Cayman Islands, Costa Rica, Cuba, Curaçao, Dominica, Dominican Republic, El Salvador, Grenada, Guatemala, Guadeloupe, Haiti, Honduras, Jamaica, Martinique, Montserrat, Nicaragua, Panama, Puerto Rico, Saint Barthelemy, Saint Kitts and Nevis, Saint Lucia, Saint Martin (French part), Saint Pierre and Miquelon, Saint Vincent and the Grenadines, Sint Maarten (Dutch part), South Georgia and the South Sandwich Islands, Trinidad and Tobago, Turks and Caicos Islands, United States Virgin Islands
Rest South America	Bolivia, Colombia, Ecuador, Falkland Islands (Malvinas), French Guiana, Guyana, Paraguay, Peru, Suriname, Uruguay, Venezuela
Former Soviet Union	Armenia, Azerbaijan, Belarus, Georgia, Kazakhstan, Kyrgyzstan, Moldova, Tajikistan, Turkmenistan, Uzbekistan
South Africa	South Africa
Ukraine	Ukraine
Egypt	Egypt
Middle East	Israel, Jordan, Lebanon, Palestine, Syria
Western Sahara, Tunisia, Morocco	Morocco, Tunisia, Western Sahara
Algeria, Libya	Algeria, Libya
Turkey	Turkey
Russia	Russia

References

1. Gazzotti, P. *et al.* Persistent inequality in economically optimal climate policies. *Nat. Commun.* **12**, 3421 (2021).
2. Leach, N. J. *et al.* Fairv2. 0.0: a generalized impulse response model for climate uncertainty and future scenario exploration. *Geosci. Model. Dev.* **14**, 3007–3036 (2021).
3. Nicholls, Z. R. *et al.* Reduced complexity model intercomparison project phase 1: introduction and evaluation of global-mean temperature response. *Geosci. Model. Dev.* **13**, 5175–5190 (2020).
4. Smith, C. J., Al Khourdajie, A., Yang, P. & Folini, D. Climate uncertainty impacts on optimal mitigation pathways and social cost of carbon. *Environ. Res. Lett.* **18**, 094024 (2023).
5. Atkinson, G., Dietz, S., Helgeson, J., Hepburn, C. & Sælen, H. Siblings, not triplets: Social preferences for risk, inequality and time in discounting climate change, economics discussion papers 2009–14. *Kiel Inst. for World Econ. http://www. economics-ejournal. org/economics/discussionpapers/2009-14* (2009).
6. Berger, L. & Emmerling, J. Welfare as equity equivalents. *J. Econ. Surv.* **34**, 727–752 (2020).
7. Adler, M. D. *Measuring social welfare: An introduction* (Oxford University Press, USA, 2019).
8. Milanovic, B. *Worlds apart: Measuring international and global inequality* (Princeton University Press, 2011).
9. Peterson, M. & Hansson, S. O. Equality and priority. *Utilitas* **17**, 299–309 (2005).
10. Temkin, L. S. Measuring inequality's badness: Does size matter? if so, how, if not, what does? *Theoria* **69**, 85–108 (2003).
11. Rabinowicz, W. The size of inequality and its badness some reflections around temkin's inequality. *Theoria* **69**, 60–84 (2003).

12. Nesje, F., Drupp, M. A., Freeman, M. C. & Groom, B. Philosophers and economists agree on climate policy paths but for different reasons. *Nat. Clim. Chang.* **13**, 515–522 (2023).

13. Drupp, M. A., Freeman, M. C., Groom, B. & Nesje, F. Discounting disentangled. *Am. Econ. Journal: Econ. Policy* **10**, 109–134 (2018).

14. Giuliani, M., Castelletti, A., Pianosi, F., Mason, E. & Reed, P. M. Curses, tradeoffs, and scalable management: Advancing evolutionary multiobjective direct policy search to improve water reservoir operations. *J. Water Resour. Plan. Manag.* **142**, 04015050 (2016).

15. Hernández-Díaz, A. G., Santana-Quintero, L. V., Coello, C. A. C. & Molina, J. Pareto-adaptive e-dominance. *Evol. computation* **15**, 493–517 (2007).