Supplements

Supplement A Supplementary figures

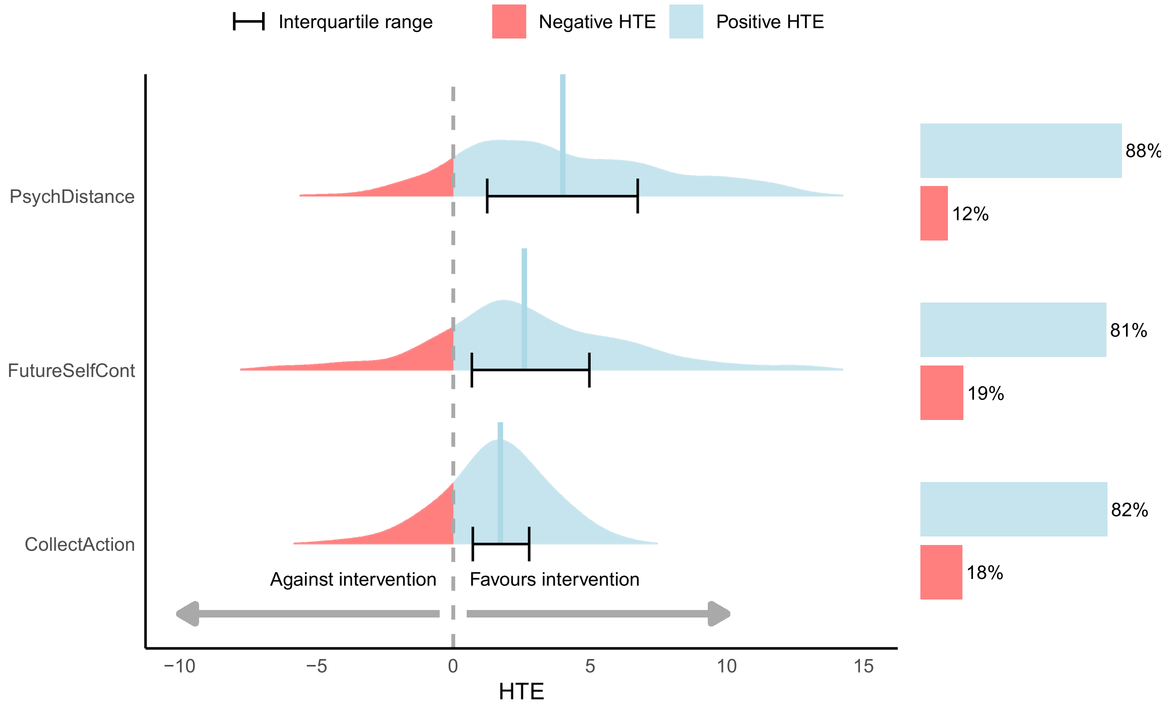
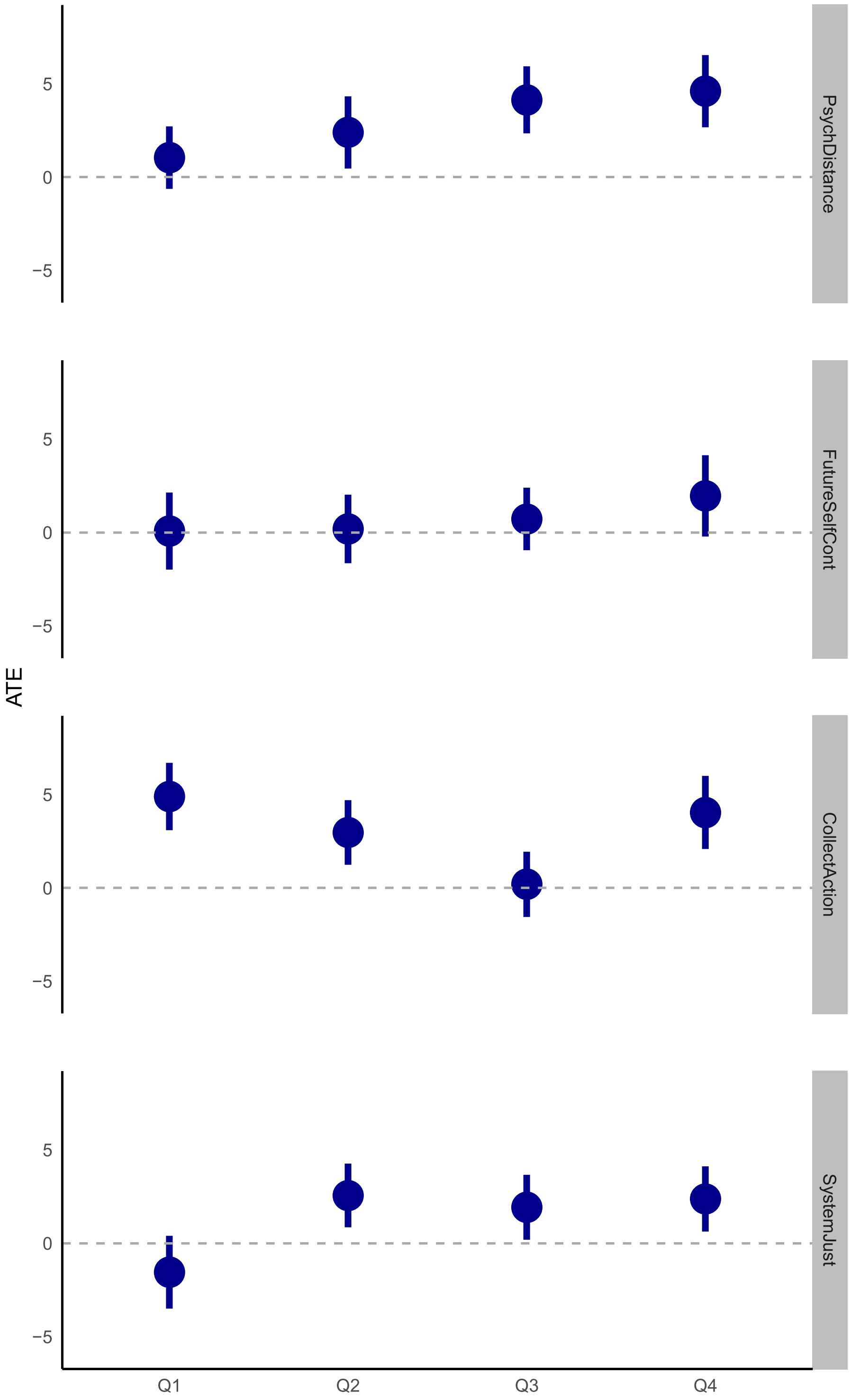
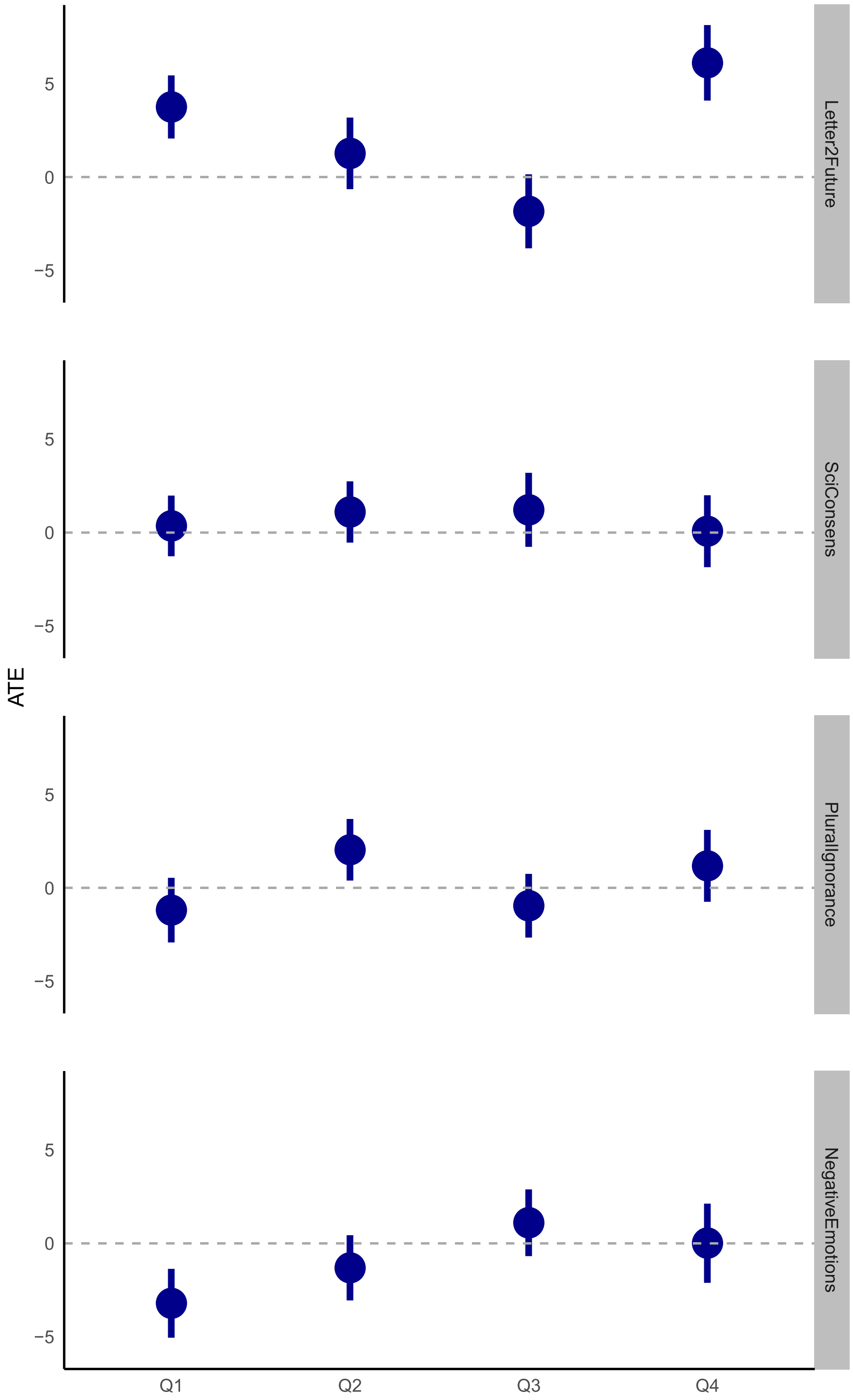


Figure S1: **Estimated HTEs across individuals and interventions from RCT 2.** The figure shows the distribution of estimated heterogeneous treatment effects (HTEs) for each of the three behavioral climate interventions, based on experimental data from RCT 2 (N = 1,610 U.S.-based participants). Each HTE represents the estimated difference in climate change belief compared to a control condition, conditional on individual variables. The bar charts indicate the proportion of individuals for whom an intervention is estimated to have a positive (blue) or negative (red) effect. For ease of interpretation, we indicate the interquartile range (horizontal whiskers) and the average treatment effect (ATE; vertical lines) estimated using a doubly robust augmented inverse propensity weighting (AIPW) method [45,46]. We note that the distributions of HTEs are expected to differ from those in Fig. 2 due to the usage of RCT 2, which included U.S.-based individuals and an extended set of variables (i.e., the attitudes and norms). For a full description of interventions, see Supplementary Table S1.


PsychDistance
FutureSelfCont
CollectAction
SystemJust
ATE
5
0
−5
Q1
Q2
Q3
Q4


ATE
5
0
−5
Letter2Future
SciConsens
PluralIgnorance
NegativeEmotions
Q1
Q2
Q3
Q4

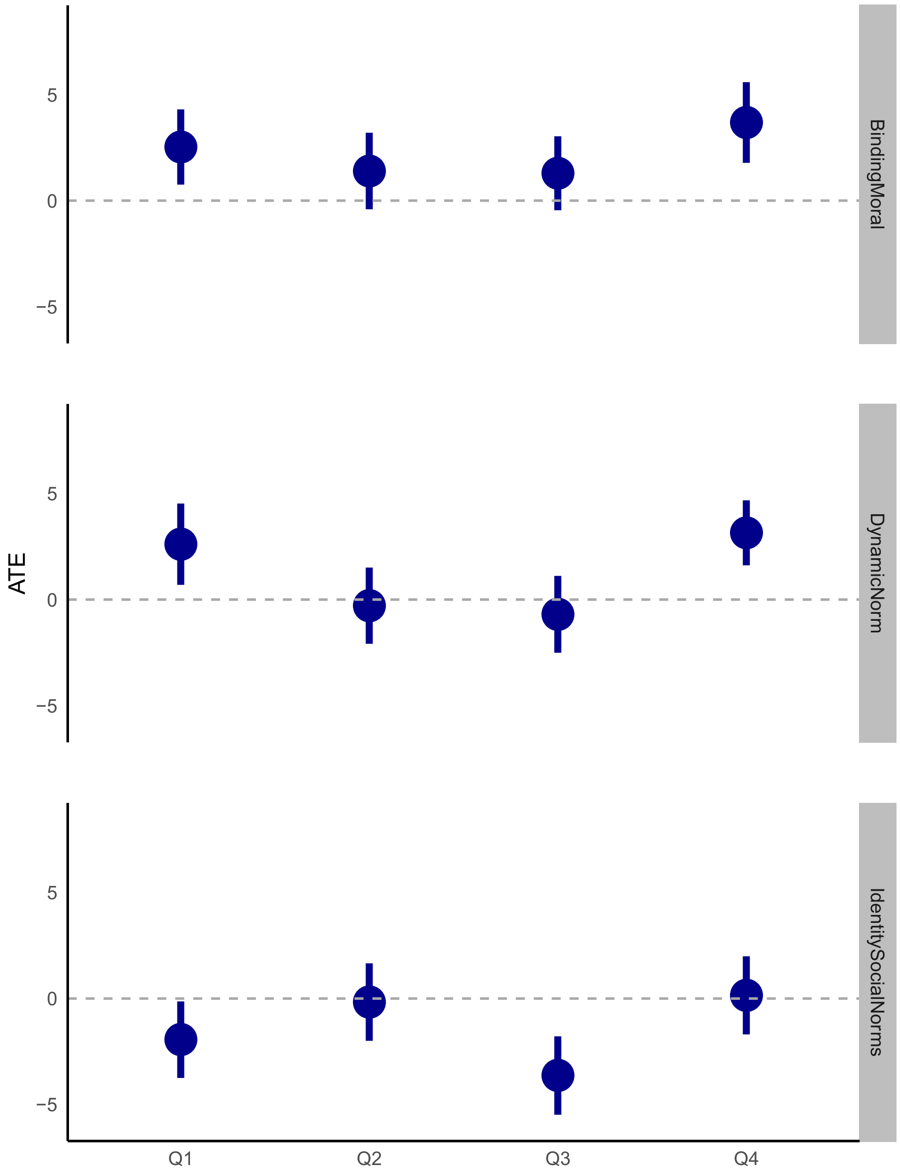


Figure S2: **External validation of the causal forest models.** The estimated HTEs for each intervention were estimated on an unseen test set, sorted in descending order, and split into quartiles. ATEs were estimated within each quartile using AIPW [45, 46]. Overall, we observe an ascending ordering of the ATEs across quartiles for each intervention. This indicates that the causal forest models are robust. The whiskers represent the SEs.

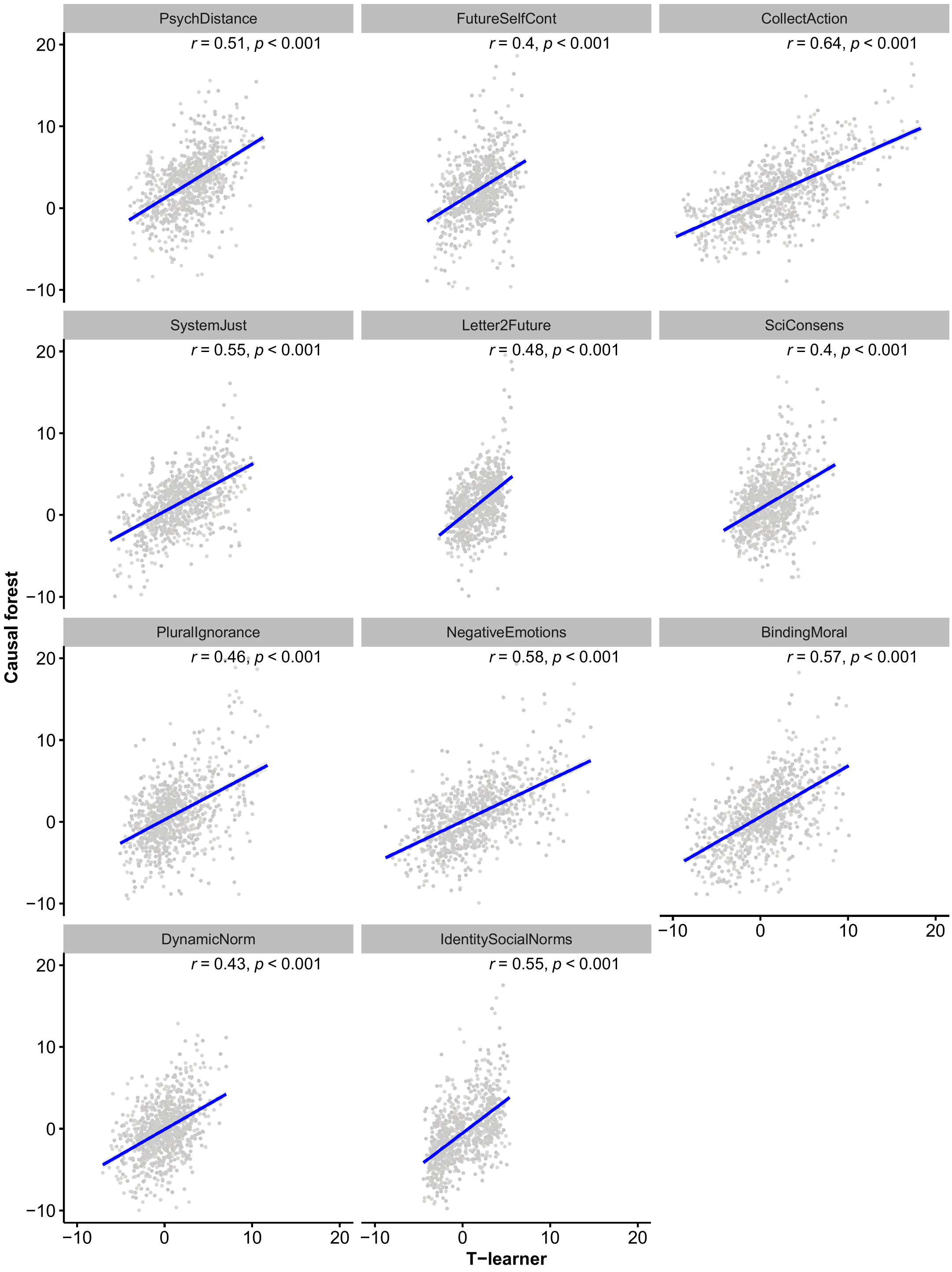


Figure S3: **Correlation of the HTE estimates obtained from the causal forest model vs. the T-learner.** As a robustness check, we estimated HTEs using the T-learner [50] (instantiated via a random forest) and compared them to the estimated HTEs from the causal forest model on the unseen test set for each intervention. We computed a linear regression line (blue) and Pearson's correlation coefficient. Overall, we found that the HTE estimates are strongly correlated (Pearson's correlation coefficient: $r > 0.4$, $p < 0.001$ for all interventions). This suggests that the HTE estimates are largely consistent across different estimation strategies.

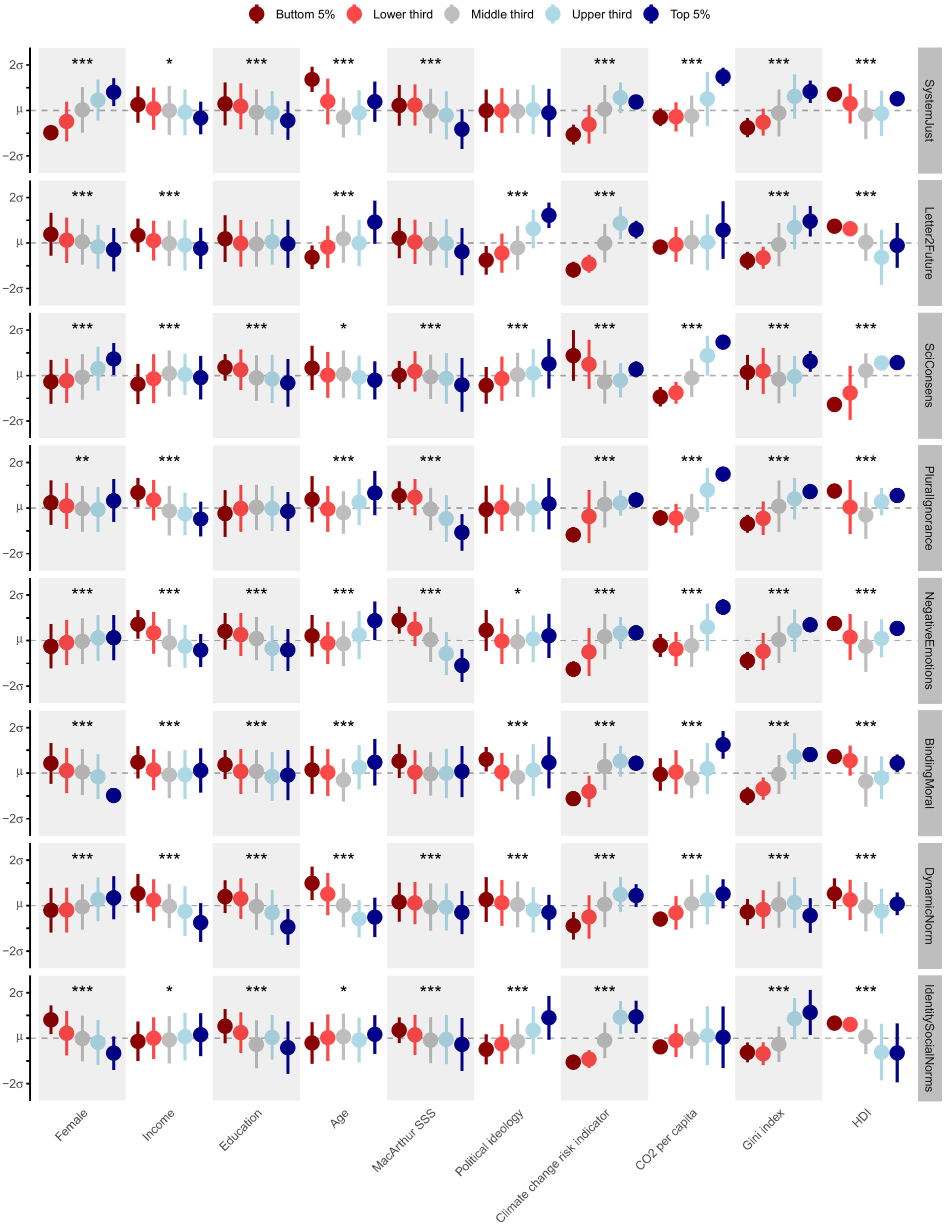


Figure S4: **Sociodemographic drivers of heterogeneity in the effectiveness of the eight other behavioral climate interventions.** Here, we assess whether there is a systematically higher intervention effectiveness for certain attitudes or norms. The plot shows how attitudes and norms, including variables referring to the environmental identity and the new ecological paradigm, are distributed across individuals. Here, we stratified the individuals into tertiles by the estimated HTEs for the eight interventions that were not covered in the main analysis. In addition to the tertiles, we also show the bottom and upper-5% tails of the estimated HTEs. Each point represents the standardized mean of a variable within a group, and the error bars show the standard error. Variables are standardized (mean = 0, SD = 1) to allow for comparability across scales. A point at $+1\sigma$ means that the average value of the variable in the specific tertile is one standard deviation higher than the overall population mean. Differences across tertiles are tested using the non-parametric Kruskal-Wallis test [53], with significance thresholds indicated as: $^{***}\,p < 0.001$, $^{**}\,p < 0.01$, $^{*}\,p < 0.05$.

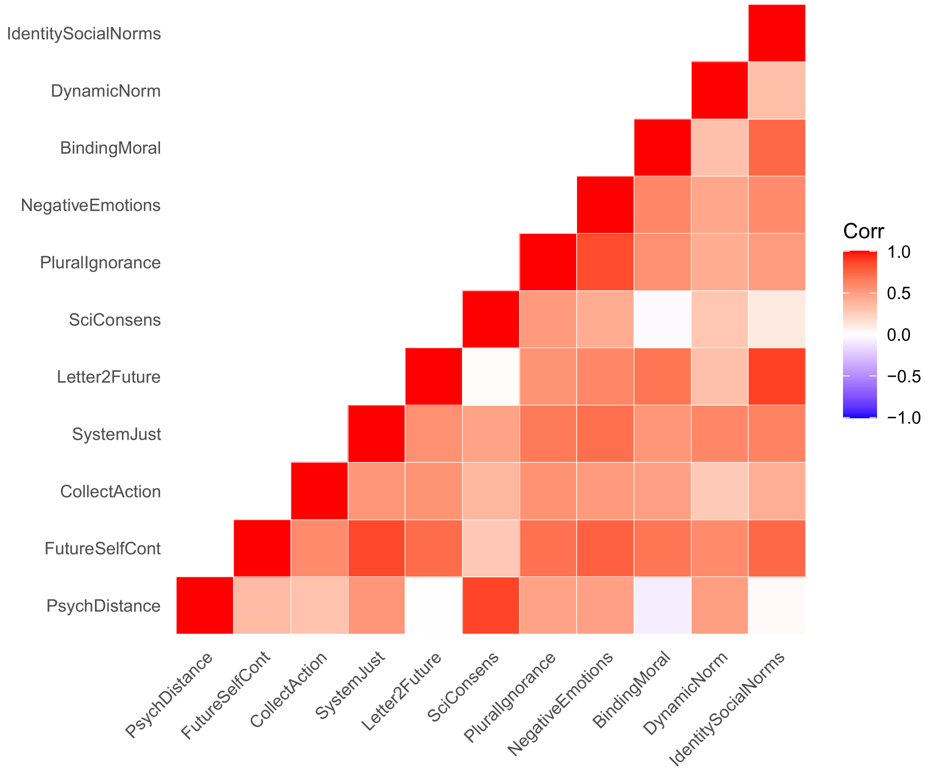


Figure S5: **Correlations between the estimated HTEs of each intervention pair.** The plot shows the correlations between the estimated HTEs of each intervention pair. Specifically, for each individual in the test set, we estimated the HTE of all eleven interventions. Then, we calculated Pearson's correlation coefficient between the estimated HTEs of each intervention pair. Overall, the HTEs of most intervention pairs are positively correlated, and a few pairs are not or even negatively correlated (e.g., PSYCHDISTANCE and BINDINGMORAL). Hence, the effectiveness of some interventions appears to be shaped by similar variables, which suggests that the effectiveness may be due to similar psychological mechanisms, while other interventions show only small correlations, implying that different psychological mechanisms are responsible for the effectiveness.

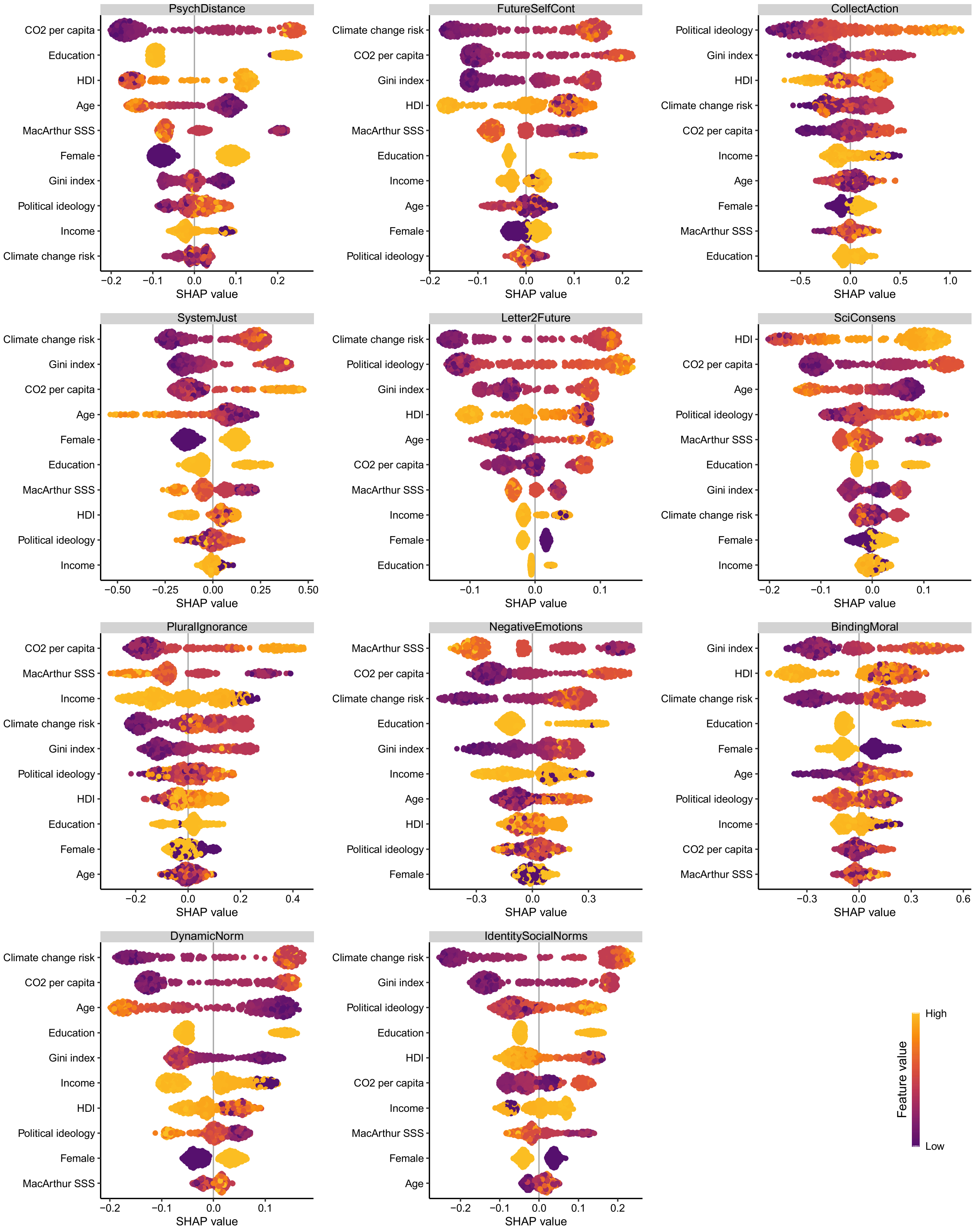


Figure S6: **SHAP value plots for all interventions in RCT 1.** Each dot represents a SHAP value [55] for a variable across different individuals, where the color indicates a high (yellow) to low (purple) SHAP value. A SHAP value of a variable measures its contribution to the estimation of an HTE relative to the average estimated HTE. Therefore, SHAP values explain drivers of heterogeneity that are associated with a larger (or smaller) HTE. Here, the variables are sorted based on their mean absolute SHAP values across the unseen test sets, so that more important determinants of HTEs are at the top.

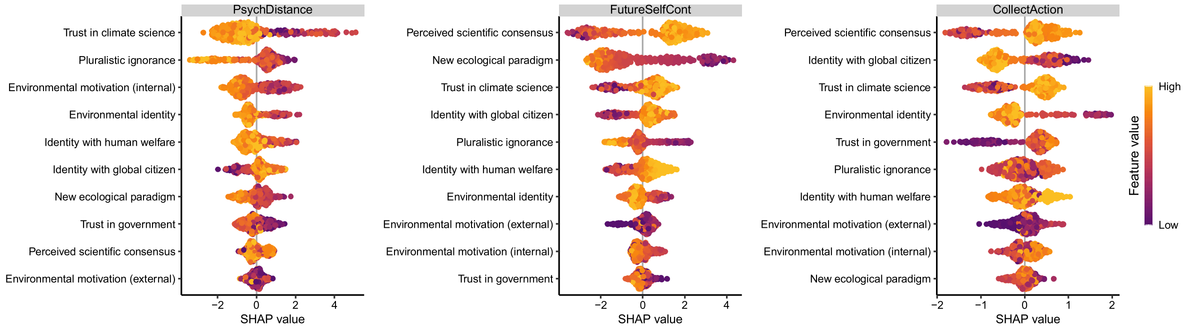


Figure S7: **SHAP value plots for all interventions in RCT 2.** Each dot represents a SHAP value [55] for a variable across different individuals, where the color indicates a high (yellow) to low (purple) SHAP value. A SHAP value of a variable measures its contribution to the estimation of an HTE relative to the average estimated HTE. Therefore, SHAP values explain drivers of heterogeneity that are associated with a larger (or smaller) HTE. Here, the variables are sorted based on their mean absolute SHAP values across the unseen test sets, so that more important determinants of HTEs are at the top.

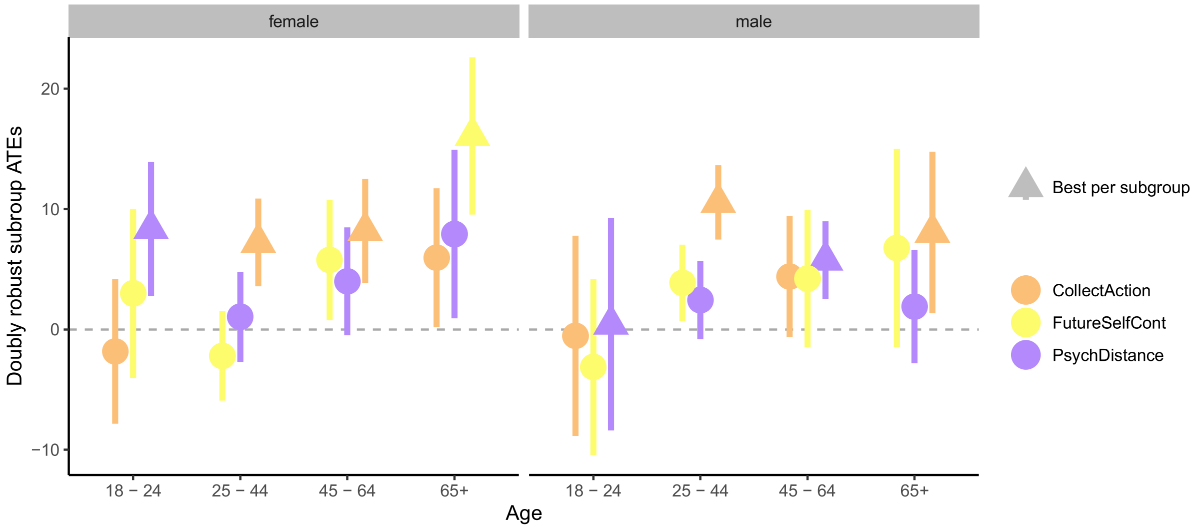
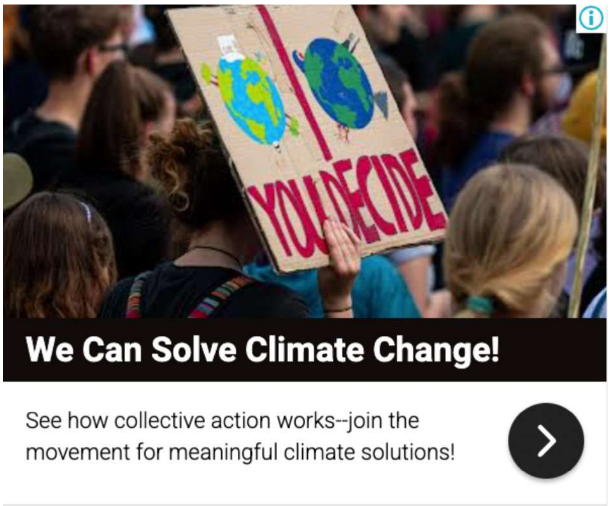


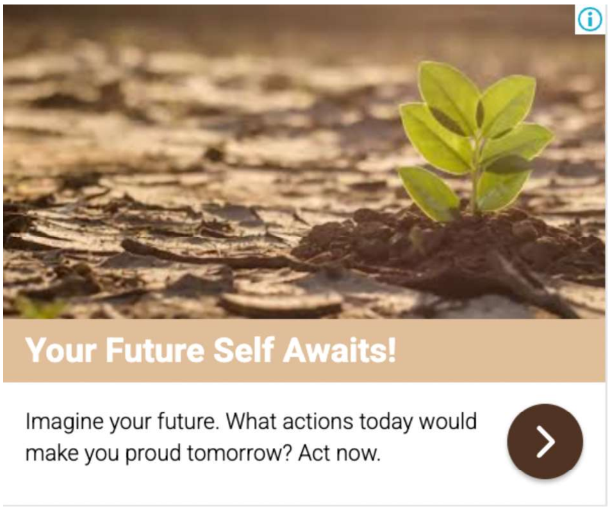
Figure S8: **ATEs for different subgroups of gender and age informing our personalized targeting approach.** Subgroup ATEs were estimated using a doubly robust AIPW method [45, 46]. The best-performing intervention per subgroup is marked as a triangle and is assigned to that subgroup within our personalized targeting approach. Hence, for example, the analysis suggests that, for females aged 18-24, the PSYCHDISTANCE intervention may be most effective, so that we allocate the PSYCHDISTANCE for this subgroup in our personalized targeting strategy.



We Can Solve Climate Change!

See how collective action works--join the movement for meaningful climate solutions!

Figure S9: **CollectAction intervention as presented in our online platform study on *Google Ads*.** The intervention was carefully constructed to resemble the same psychological mechanisms as presented in [18].



Your Future Self Awaits!

Imagine your future. What actions today would make you proud tomorrow? Act now.

Figure S10: **FutureSelfCont intervention as presented in our online platform study on *Google Ads*.** The intervention was carefully constructed to resemble the same psychological mechanisms as presented in [18].

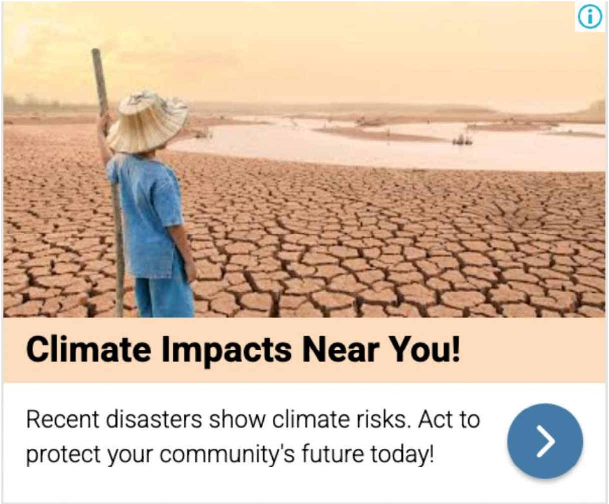

Climate Impacts Near You!
Recent disasters show climate risks. Act to protect your community's future today!

Figure S11: **PsychDistance intervention as presented in our online platform study on *Google Ads*.** The intervention was carefully constructed to resemble the same psychological mechanisms as presented in [18].

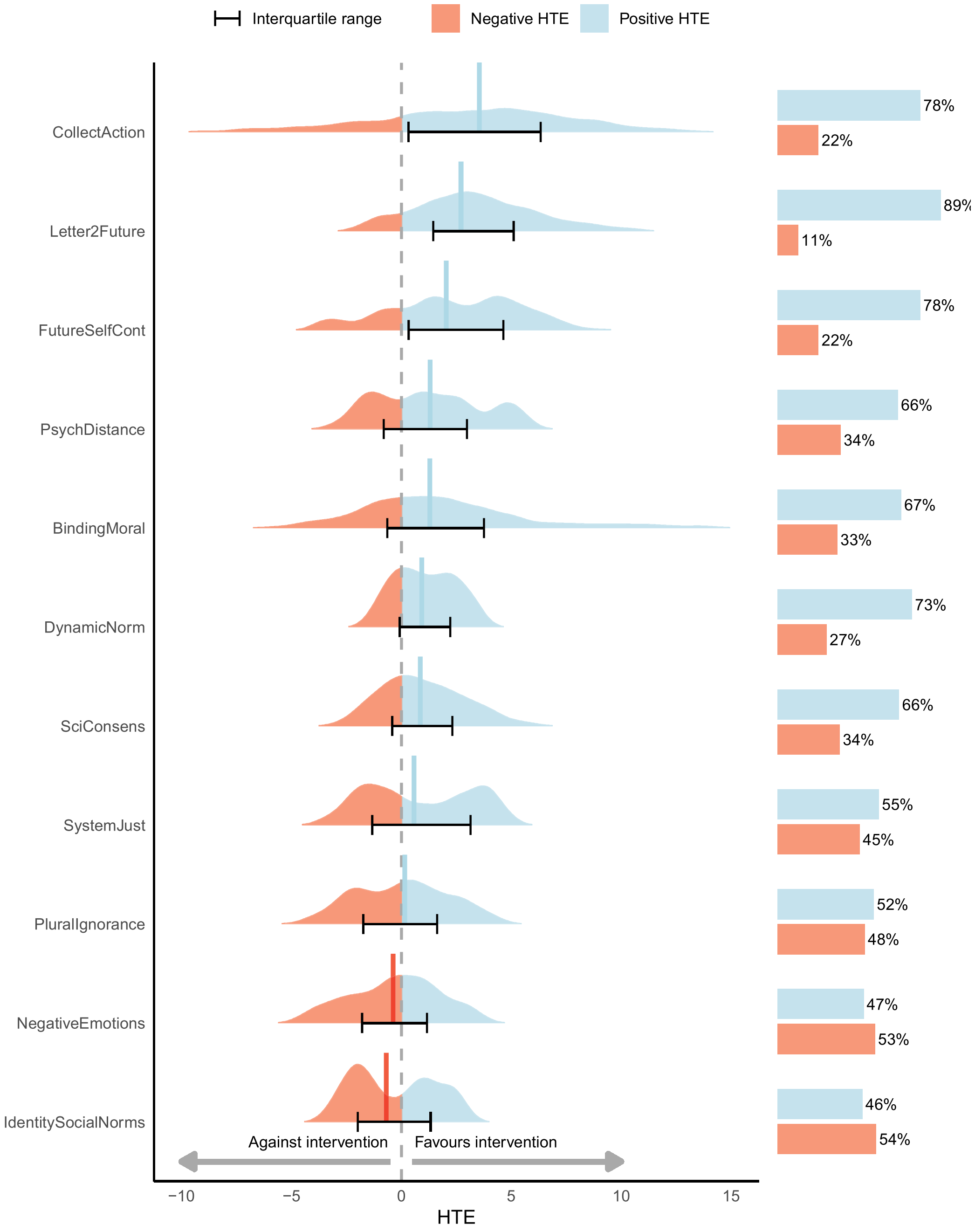


Figure S12: **Estimated HTEs across individuals and interventions on the outcome climate policy support.** The figure shows the distribution of estimated heterogeneous treatment effects (HTEs) on the outcome climate policy support for each of the 11 behavioral climate interventions, based on experimental data from RCT 1 (N = 59,508 participants across 63 countries). Each HTE represents the estimated difference in climate change belief compared to a control condition, conditional on individual variables. The bar charts indicate the proportion of individuals for whom an intervention is estimated to have a positive (blue) or negative (red) effect. For ease of interpretation, we indicate the interquartile range (horizontal lines) and the average treatment effect (ATE) estimated using a doubly robust augmented inverse propensity weighting (AIPW) method [45,46]. For a full description of interventions, see Supplementary Table S1.

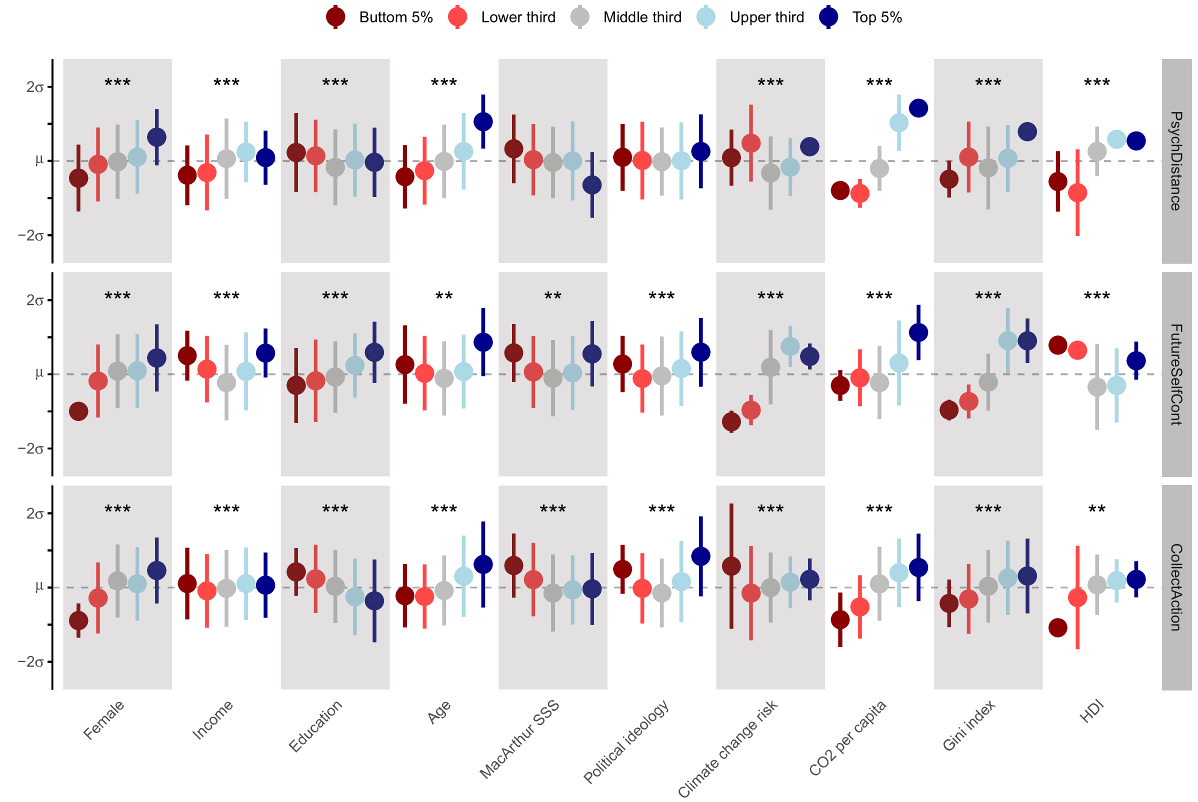


Figure S13: **Sociodemographic drivers of heterogeneity in the effectiveness of behavioral climate interventions on the outcome climate policy support.** Here, we assess whether certain characteristics are systematically associated with higher intervention effectiveness on climate policy support. The plot shows how the sociodemographic variables, including the MacArthur scale of subjective social status (MacArthur SSS) and the human development index (HDI), are distributed across individuals with different estimated effects of behavioral climate interventions. Here, we stratified the individuals into tertiles by the estimated HTEs for the three selected interventions: PSYCHDISTANCE, FUTURESELFCONT, and COLLECTACTION. In addition to the tertiles, we also show the bottom and upper-5% tails of the estimated HTEs. Each point represents the standardized mean of a variable within a group, and the error bars show the standard error. Variables are standardized (mean = 0, SD = 1) to allow for comparability across scales. Hence, a point at $+1\sigma$ means that the average value of the variable in the specific tertile is one standard deviation higher than the overall population mean. Differences across tertiles are tested using the non-parametric Kruskal-Wallis test [53], with significance thresholds indicated as: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

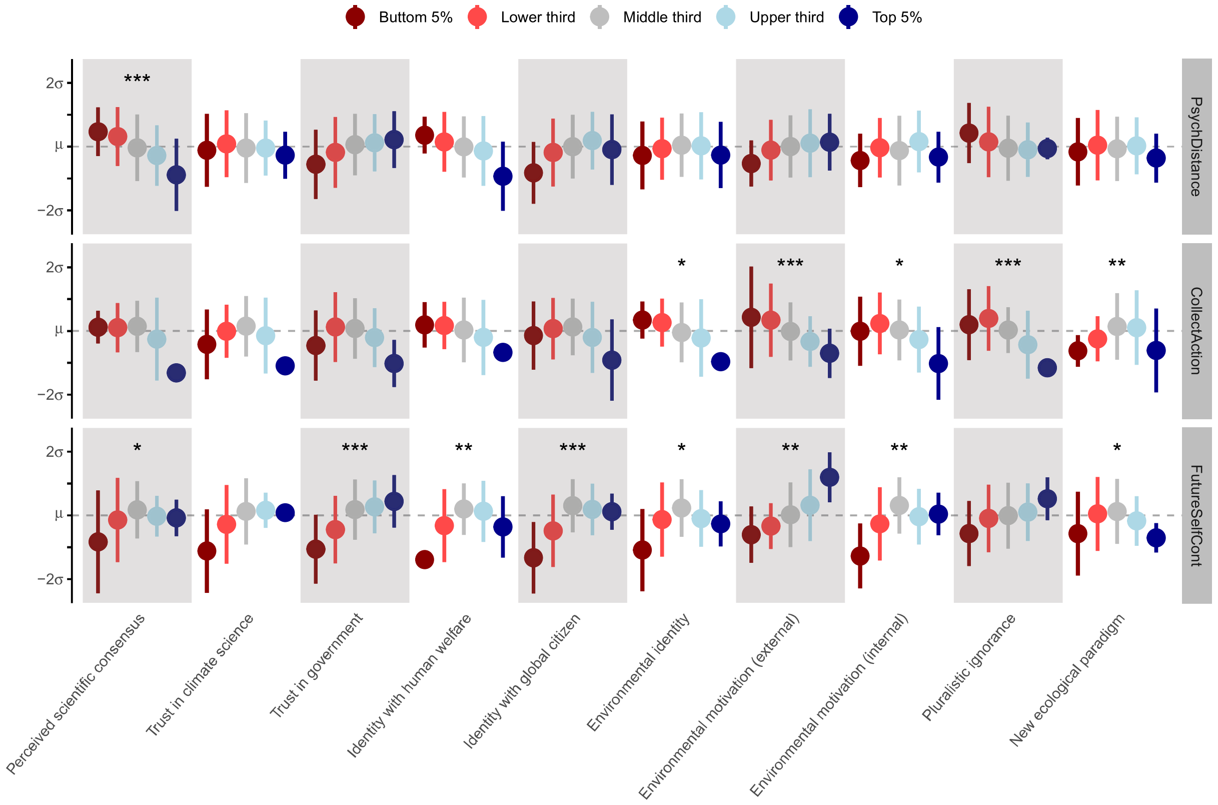


Figure S14: **Attitudinal and normative drivers of heterogeneity in the effectiveness of behavioral climate interventions on the outcome climate policy support.** Here, we assess whether certain characteristics are systematically associated with higher intervention effectiveness on climate policy support. The plot shows how attitudes and norms are distributed across individuals with different estimated effects of behavioral climate interventions. Here, we stratified the individuals into tertiles by the estimated HTEs for the three interventions: PsychDistance, FutureSelfCont, and CollectAction. In addition to the tertiles, we also show the bottom and upper-5% tails of the estimated HTEs. Each point represents the standardized mean of a variable within a group, and the error bars show the standard error. Variables are standardized (mean = 0, SD = 1) to allow for comparability across scales. Hence, a point at $+1\sigma$ means that the average value of the variable in the specific tertile is one standard deviation higher than the overall population mean. Differences across tertiles are tested using the non-parametric Kruskal-Wallis test [53], with significance thresholds indicated as: $^{***}\,p < 0.001$, $^{**}\,p < 0.01$, $^{*}\,p < 0.05$.

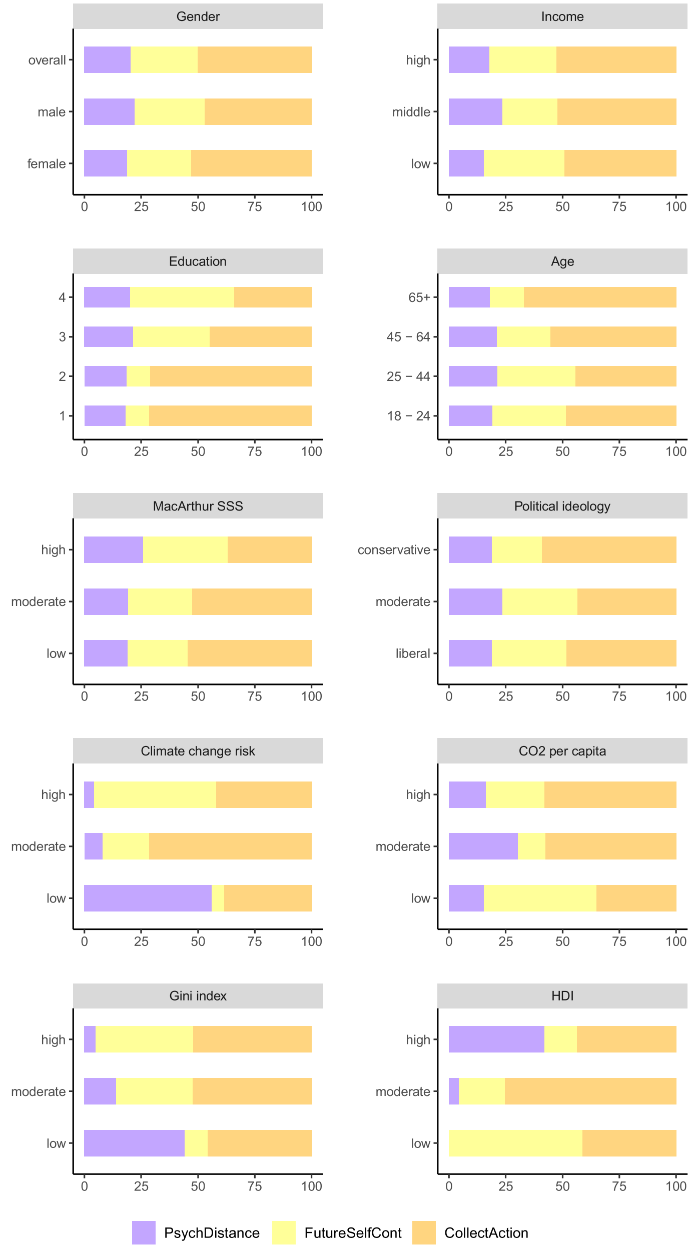


Figure S15: **Optimal interventions by sociodemographic subgroups using the outcome climate policy support.** The proportions of how often each intervention is optimal (x-axis, from 0 to 100%) and therefore has the largest HTE within that subgroup as determined by the causal forest model. In other words, the optimal intervention is defined as the one with the highest HTE on climate policy support. Education levels are encoded using a discrete categorization (i.e., 1: up to grade school, 2: up to high school, 3: college/undergraduate degree/certificate training, 4: doctorate degree, medical degree, etc.). We categorize the other continuous variables into three equally sized subgroups (e.g., high, middle, and low-level income). We denote the MacArthur scale of subjective social status as MacArthur SSS and the human development index as HDI. For a full description of the sociodemographics, see Supplementary Table S2.

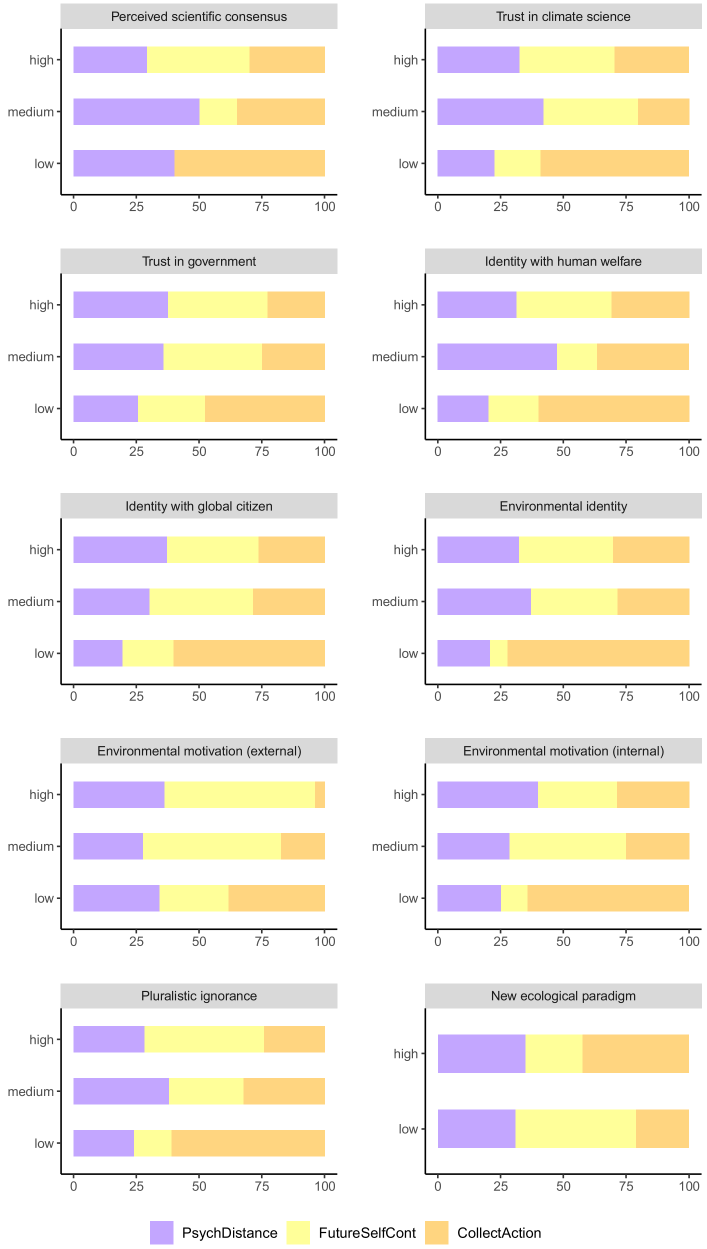


Figure S16: **Optimal interventions by attitudinal and normative subgroups using the outcome climate policy support.** The proportions of how often each intervention is optimal (x-axis, from 0 to 100%) and therefore has the largest HTE within that subgroup as determined by the causal forest model. In other words, the optimal intervention is defined as the one with the highest HTE on climate policy support. We categorize the continuous variables into three equally sized subgroups (e.g., high, middle, and low-level environmental identity). For a full description of the sociodemographics, see Supplementary Table S3.

Supplement B Supplementary tables

Intervention	**Description**
PsychDistance	Frames climate change as a proximal risk using examples of recent natural disasters caused by climate change in each participant's nation and prompts them to write about the climate impacts on their community.
FutureSelfCont	Emphasizes the future self-continuity by asking each participant to project themselves into the future and write a letter addressed to themselves in the present, describing the actions they would have wanted to take regarding climate change.
CollectAction	Features examples of successful collective action that have had meaningful effects on climate policies (e.g., protests) or have solved past global issues (e.g., the restoration of the ozone layer).
SystemJust	Frames climate change as threatening to the way of life of each participant's nation and makes an appeal to climate action as the patriotic response.
Letter2Future	Emphasizes how one's current actions affect future generations by asking participants to write a letter to a socially close child who will read it in 25 years when they are an adult, describing current actions toward ensuring a habitable planet.
SciConsens	Informs participants that "99% of expert climate scientists agree that Earth is warming and climate change is happening, mainly because of human activity."
PluralIgnorance	Presents real public opinion data collected by the United Nations that show what percentage of people in each participant's country agree that climate change is a global emergency.
NegativeEmotions	Exposes participants to ecologically valid scientific facts regarding the impacts of climate change framed in a "doom and gloom" style of messaging that were drawn from different real-world news and media sources.
BindingMoral	Invokes authority (e.g., "from scientists to experts in the military, there is near universal agreement"), purity (e.g., keep our air, water, and land pure), and ingroup-loyalty (e.g., "it is the American solution") moral foundations.
DynamicNorm	Informs participants of how country-level norms are changing and "more and more people are becoming concerned about climate change," suggesting that people should take action.
IdentitySocialNorms	Combines referencing a social norm ("a majority of people are taking steps to reduce their carbon footprint") with an invitation to "join in" and work together with fellow citizens toward this common goal.

Table S1: **Descriptions of the 11 behavioral climate interventions.** The behavioral climate interventions were developed by behavioral scientists and pre-screened for cross-cultural feasibility and relevance to climate attitudes [18].

Variable	Description
Individual-level sociodemographic variables	
Gender	*What is your gender?*; response options: Male, Female, Prefer not to say, Non-binary/third gender/other.
Income	*What is your total yearly family/household income?*; reported using 8 ordinal brackets.
Education	*How many years of formal education have you completed?*; categorized into: 0-6 (grade school), 7-12 (high school), 13-16 (college/undergraduate), 17+ (doctorate/professional degrees).
Age	*How old are you?*; single-line numerical entry (in years).
MacArthur scale of subjective social status	*Please choose the rung where you think you stand at this time in your life relative to other people in your country*; MacArthur Scale from 1 (lowest) to 10 (highest).
Political orientation	*What is your political orientation for social issues (e.g., health care, education, etc.) and economic issues (e.g., taxes)?*; each item rated on a 0-100 scale. Items were averaged into a composite conservatism score (α = 0.84, see [18]).
Country-level sociodemographic variables	
Climate change risk indicator	Index measuring national exposure to climate-related disasters; sourced from World Bank (https://prosperitydata360.worldbank.org/en/dataset/IMF+RDD). Higher values indicate higher risk.
CO_2 per capita	Production-based CO_2 emissions per person (measured in tons); sourced from https://ourworldindata.org/co2-emissions#per-capita-co2-emissions.
GINI coefficient	Inequality index measuring national income distribution; sourced from World Bank (https://data.worldbank.org/indicator/SI.POV.GINI). Higher values indicate greater inequality.
HDI	Human development index combining life expectancy, education, and per capita income; sourced from https://ourworldindata.org/human-development-index.

Table S2: **Sociodemographic variables used as predictors of heterogeneity in our analysis.**

Variable	Description
Perceived scientific consensus	*“To the best of your knowledge, what percentage of climate scientists have concluded that human-caused climate change is happening?”* Measured on a scale from 0 to 100.
Trust in climate science	*“On average, how competent are climate change research scientists?”*; *“How much do you trust scientific research about climate change?”*. Measured on a scale from 0 to 100.
Trust in government	*“On average, how much do you trust your government?”* Measured on a scale from 0 to 100.
Identity with human welfare	*“To what degree do you see yourself as someone who cares about human welfare?”*. Measured on a scale from 0 to 100.
Identity with global citizens	*“To what degree do you think of yourself as a global citizen?”*. Measured on a scale from 0 to 100.
Environmental identity	*“To what degree do you see yourself as someone who cares about the natural environment?”*; *“To what degree are you pleased to be someone who cares about the natural environment?”*; *“To what degree do you feel strong ties with others who care about the natural environment?”*; *“To what degree do you identify with others who care about the natural environment?”*. Measured on a scale from 0 (not at all) to 100 (very much so) and finally averaged.
External environmental motivation	*“Because of today’s politically correct standards, I try to appear pro-environmental.”*; *“I try to hide my negative thoughts about pro-environmental behavior in order to avoid negative reactions from others.”*; *“If I acted anti-environmental, I would be concerned that others would be angry with me.”*; *“I attempt to appear pro-environmental in order to avoid disapproval from others.”*; *“I try to act pro-environmental because of pressure from others.”*. Measured on a scale from 0 (strongly disagree) to 100 (strongly agree) and finally averaged.
Internal environmental motivation	*“I attempt to behave pro-environmentally because it is personally important to me.”*; *“According to my personal values, acting non-environmental is OK.”*; *“I am personally motivated by my beliefs to be pro-environmental.”*; *“Because of my personal values, I believe that acting anti-environmental is wrong.”*; *“Being pro-environmental is important to my self-concept.”*. Measured on a scale from 0 (strongly disagree) to 100 (strongly agree) and finally averaged.
Pluralistic ignorance	*“What percentage of people in your country do you think would agree with the statement: Climate change is a global emergency?”* Measured on a scale from 0 to 100.
New Ecological Paradigm (NEP)	Fifteen items assessing general environmental beliefs (e.g., views on human-nature relationships), based on the NEP scale [72]. Widely used to measure environmental attitudes [73]. The items are finally averaged.

Table S3: **Attitudinal and normative variables used as predictors of heterogeneity in our analysis**

	Dependent variable: Click								
	Overall	Females (18-24)	Females (25-44)	Females (45-64)	Females (65+)	Males (18-24)	Males (25-44)	Males (45-64)	Males (65+)
Personalized	0.273^{***}	−0.044	0.113^{*}	0.260^{***}	0.125^{*}	0.168^{**}	0.127^{*}	−0.019	0.820^{***}
	(0.020)	(0.057)	(0.057)	(0.057)	(0.055)	(0.056)	(0.059)	(0.057)	(0.058)
Constant	-4.455^{***}	-3.460^{***}	-3.813^{***}	-4.651^{***}	-5.243^{***}	-3.490^{***}	-3.474^{***}	-4.073^{***}	-5.237^{***}
	(0.014)	(0.040)	(0.040)	(0.040)	(0.039)	(0.040)	(0.041)	(0.041)	(0.040)
Observations	772,777	43,125	55,033	115,621	237,947	40,559	37,756	74,240	168,496
Log likelihood	−53,538.160	−5,783.487	−5,979.768	−6,810.496	−8,206.704	−5,747.006	−5,327.224	−6,267.137	−7,120.190
AIC	107,080.300	11,570.970	11,963.540	13,624.990	16,417.410	11,498.010	10,658.450	12,538.270	14,244.380

Note: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

Table S4: **Logistic regression results for the external validation of our personalized targeting strategy on *Google Ads*.** Here, we report the overall and subgroup-specific coefficient estimates from our logistic regressions.

Supplement C Related work and theoretical mechanisms

For our main study, we used the three top-performing interventions from [18]. These interventions were: (1) PSYCHDISTANCE, (2) FUTURESELFCONT, and (3) COLLECTACTION. In the following, we provide related work and theoretical explanations for each of these interventions.

The PSYCHDISTANCE intervention [18] frames climate change as a local, immediate, and personally relevant threat. People frequently perceive climate change as psychologically distant—e.g., in space, time, or social relevance—which reduces their urgency of acting [74]. According to *Construal Level Theory*, making climate impacts seem closer, immediate, and more personal increases concrete thinking and, thus, the likelihood of acting [75]. Empirical studies show that localizing and personalizing climate threats significantly increases public concern and willingness to act [76, 77].

The COLLECTACTION intervention [18] focuses on successful climate movements to foster collective efficacy. Collective efficacy—i.e., the belief in a group's capacity to achieve desired outcomes—strongly predicts engagement in collective actions [78,79]. Research shows that awareness of past successful movements enhances people's sense of efficacy and motivates participation in collective climate action [80].

Lastly, the FUTURESELFCONT intervention [18] enhances climate action by increasing the psychological connection to one's future self. People often neglect distant future outcomes because they perceive their future self as a different person [81, 82]. Strengthening continuity with one's future self motivates more future-oriented decisions [83]. Empirical evidence confirms that vivid perspective-taking exercises, such as writing letters from one's future self, significantly increase future-minded behaviors [84]. As such, connecting present behavior directly to personal future consequences fosters responsibility and proactive engagement in climate protection.

Supplement D Background on HTE estimation

Estimand

We adopt the potential outcomes framework [85, 86] to define our causal quantities of interest, namely, the heterogeneous treatment effect (HTE). The potential outcomes framework provides a principled basis for causal inference by comparing the outcomes that an individual would experience under different treatment conditions $t \in \{0, 1, \dots, K\}$, where $t = 0$ denotes the control condition and $K = 11$ corresponds to the different behavioral interventions tested. For each individual i, we define the potential outcome under treatment condition t as $Y_i(t)$. Due to the fundamental problem of causal inference, only one of all potential outcomes is observed for each individual, depending on the assigned treatment condition [85, 87].

Formally, the HTE for intervention t is defined conditional on observed variables X_i, i.e.,

$$\mathrm{HTE}_t(X_i) = \mathbb{E}[Y_i(t) - Y_i(0) \mid X_i], \text{ for } t \in \{1, 2, \dots, K\}. \tag{S1}$$

The HTE thus captures between-subject variability in the intervention effectiveness and thus allows for a more granular understanding of how effects vary across individuals. This is unlike the ATE, which, for intervention t, is defined as

$$\mathrm{ATE}_t = \mathbb{E}[Y_i(t) - Y_i(0)], \text{ for } t \in \{1, 2, \dots, K\}. \tag{S2}$$

Hence, the ATE provides a single "population-wide" estimate of the treatment effect and thus obscures potential heterogeneity in the intervention effectiveness across individuals.

Our analysis uses experimental data in which individuals were randomly assigned to one of the expert-crowdsourced behavioral climate interventions or a control condition. As a result, the assumptions required for valid causal inference—the stable unit treatment value assumption

(SUTVA), positivity, and unconfoundedness—are satisfied by design [85, 86]. This eliminates complex adjustment procedures and ensures identification of the causal estimand and thus adds to the internal validity of the treatment effect estimates.

D.1 Procedure

To estimate HTEs, we applied the causal forest model [38, 39], a nonparametric causal machine learning method designed to capture heterogeneity in treatment effects. Causal forests partition the variable space to capture complex, high-dimensional interactions within the HTEs. In our setup with 11 different treatments, we estimate separate causal forests for each intervention $t \in \{1, 2, \ldots, K\}$, each relative to the control group $t = 0$.

To improve the robustness and efficiency of HTE estimation, we apply orthogonalization [88] to our data. Specifically, we leverage the known propensity scores from the random assignment for each $t \in \{0, 1, \ldots, K\}$ (i.e., $e_t(X_i) = 0.5$) and estimate the nuisance outcome function $\hat{m}_t(X_i) = \mathbb{E}[Y_i \mid X_i]$ using a regression forest. This orthogonalization step has several advantages, even in the presence of randomization. First, the climate change belief outcome is non-normally distributed and bounded (e.g., on a Likert or probability scale), making standard parametric assumptions inappropriate [18]. Orthogonalization allows for the use of flexible, nonparametric machine learning methods to model complex relationships in the outcome model. In addition, by partialling out variation from direct contributions of the variables on climate change belief before estimating HTEs, orthogonalization leads to more stable and lower-variance estimates of the HTEs. We then employ the causal forest model to optimize the following objective function based on the orthogonalization:

$$\widehat{\mathrm{HTE}}_t(X) = \arg \min_{\mathrm{HTE}_t(X)} \hat{\mathbb{E}} \left[\left((Y - \hat{m}_t(X_i)) - HTE_t(X) \cdot \left(\mathbb{1}_{\{T_i = t\}} - e_t(X_i) \right) \right)^2 \right]. \qquad \text{(S3)}$$

D.2 Comparison to ATEs

To estimate ATEs, we use the augmented inverse propensity weighted (AIPW) estimator [45,46].[1] This nonparametric estimator was chosen for several reasons. First, it offers flexibility in capturing nonlinear relationships and interactions between variables and outcomes. Second, it is robust to violations of parametric assumptions, such as skewness, non-normality, and bounded outcomes, which are present in our underlying data [18]. Third, AIPW is a doubly robust estimator, meaning it yields consistent estimates if either the outcome model or the propensity score model is correctly specified. In our case, the known propensity score from random assignment (i.e., 0.5) guarantees this condition is satisfied. Finally, by incorporating variables through a regression forest in the outcome model, AIPW reduces the variance of ATE estimates compared to simple difference-in-means, even in randomized settings.

The AIPW estimator consists of three components: (i) the propensity score $e(X_i)$ is known here given the randomization in our studies, i.e., $e(X_i) = 0.5$, (ii) the outcome model for the individuals in the control condition is defined as $\hat{m}_0(X_i) = E[Y_i \mid T_i = 0, X_i]$, and (iii) the outcome model of the individuals in one of the interventions $t \in \{1, 2, \ldots, K\}$ defined as $\hat{m}_t(X_i) = E[Y_i \mid T_i = t, X_i]$. Using the three components, the ATE for intervention $t \in \{1, 2, \ldots, K\}$ can be calculated using the AIPW estimator via

$$\widehat{\text{ATE}}_{\text{AIPW,t}} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\mathbb{1}_{\{T_i=t\}}Y_i}{e(X_i)} - \frac{\mathbb{1}_{\{T_i=0\}}Y_i}{1-e(X_i)} + \left(1 - \frac{\mathbb{1}_{\{T_i=t\}}}{e(X_i)}\right)\hat{m}_t(X_i) - \left(1 - \frac{\mathbb{1}_{\{T_i=0\}}}{1-e(X_i)}\right)\hat{m}_0(X_i)\right]. \quad \text{(S4)}$$

Implementation details

We used the implementation of the causal forest from `grf` (version 2.3.2) [67]. We used 3-fold cross-validation to optimize the hyperparameters of the causal forests. This is done by randomly sampling sets of hyperparameter values, training small forests of size 500 for each, and evaluating their performance using the out-of-bag error. The optimal parameters are then selected by minimizing a smoothed version of this error across sampled configurations. The hyperparameters and

[1] We apply the AIPW estimator for both overall ATEs and subgroup ATEs throughout our analyses.

their value spaces are given in Table S5. For all outcome models $\hat{m}_t(X_i)$ with $t \in \{0, 1, \dots, K\}$ we used random forests [65] based on the implementation in `grf` [67]. We used the hyperparameters in Table S6 for the random forests.

Hyperparameter	Sampled range (from uniform $u \sim \mathcal{U}(0,1)$)
Minimum node size	$2^{u \cdot (\log_2(n) - 4)}$
Sample fraction	$0.05 + 0.45 \cdot u$
Number of variables	$\lceil u \cdot \min(p, \sqrt{p} + 20) \rceil$
Maximum imbalance of split	$u/4$
Imbalance penalty	$-\log(u)$
Honesty fraction	$0.5 + 0.3 \cdot u$
Honesty prune leaves	`TRUE` if $u < 0.5$, else `FALSE`
Number of trees	4000

Note: p refers to the number of variables

Table S5: **Hyperparameters for our causal forests.**

Hyperparameter	Value
Number of trees	2000
Number of variables	$\sqrt{p} + 20$
Minimum node size	5
Maximum imbalance of split	0.05

Note: p refers to the number of variables

Table S6: **Hyperparameters for random forests used for our outcome models.**