Machine learning reveals how personalized climate communication can both succeed and backfire

Totte Harinen
Toyota Research Institute
totte.harinen@tri.global

Alexandre Filipowicz
Toyota Research Institute
alex.filipowicz@tri.global

Shabnam Hakimi
Toyota Research Institute
shabnam.hakimi@tri.global

Rumen Iliev
Toyota Research Institute
rumen.iliev@tri.global

Matthew Klenk
Toyota Research Institute
matt.klenk@tri.global

Emily Sarah Sumner
Toyota Research Institute
emily.sumner@tri.global

Abstract

Different advertising messages work for different people. Machine learning can be an effective way to personalise climate communications. In this paper we use machine learning to reanalyse findings from a recent study, showing that online advertisements increased climate change belief in some people while resulting in decreased belief in others. In particular, we show that the effect of the advertisements could change depending on a person's age and ethnicity. Our findings have broad methodological and practical applications.

Social and behavioral sciences are increasingly making use of advertising campaigns to help change beliefs about topics such as climate change and vaccine hesitancy [7, 15]. However, advertisement campaigns do not always have the same effect on each viewer and, if not carefully understood, can *backfire* [22, 15]. For example, campaigns designed to decrease vaccine hesitancy have inadvertently made some segments of the population *less* likely to get vaccinated [15]. These backfire effects are well understood in domains such as personalized advertisement, and recent machine learning research has focused on measuring heterogeneous treatment effects (HTEs). This class of analyses aims to distinguish people for whom a campaign might work from those for whom it might backfire [2, 24, 11, 20, 2, 24, 12, 9, 1, 26, 27, 10, 19].

Despite their recent popularity in machine learning, HTE analysis methods have been slow to permeate social and behavioral sciences [3]. This slow trend has been recently been flagged as a considerable barrier for behavioral science, as the methods traditionally used to evaluate large scale interventions largely ignore treatment heterogeneity across different population segments [3]. As a consequence, the insights gleaned from these interventions are likely limited in their utility and generalizability.

In this article we provide an example of how HTEs can provide deeper insights into the varied effects of behavioral interventions of different population segments. We reanalyse the results of a field experiment conducted in the US that used personalised ad campaigns to change beliefs about climate change [7]. The ad campaign was specifically designed to influence beliefs in Republicans, a political group that has historically expressed skepticism in climate change. To target Republicans, the ads centered around topics generally deemed important by Republicans (e.g. faith, security and immigration). Because the campaign was specifically targeted to individuals identifying as Republican, the authors analyzed the effect of their intervention on people from different political

NeurIPS 2022Workshop on Causal Machine Learning for Real-World Impact (CML4Impact 2022).

affiliations. They found that their intervention successfully increased Republican beliefs in climate change by 7%, whereas beliefs from respondents with other political affiliations remained unaffected. Although encouraging, by focusing solely on political affiliation, the authors failed to examine the extent to which their intervention affected other population segments.

Here we use machine learning based HTE analyses to show that while the ads did reliably increase beliefs in climate change for some segments, they also reliably *reduced* beliefs in climate change in others, in some cases by as much as 45%. Our analysis highlights the importance of proper HTE analyses in the interpretation and recommendations made from interventions both to understand for whom an intervention might be effective, but also importantly for whom the intervention might backfire.

Methods

We use a causal machine learning approach to reanalyze the treatment effect of the intervention conducted in [7]. We focus on the difference in the belief that climate change is happening between individuals in the treatment group (who were targeted with ads) and individuals in the control group (who were not targeted). This analysis identifies both "positive" and "negative" treatment effects, corresponding to increases and decreases, respectively, in beliefs in climate change. Here we use machine learning to predict how this treatment effect changes for different groups of individuals. As predictors, we use all 11 demographic variables collected by [7] (Figure 2).

The outcome variable in our analysis is people's answer to the question "Do you think that global warming is happening?" The possible answers to this question were "Yes", "Don't know" and "No". Following the original analysis in [7], we binarise the outcome such that "Yes" is recoded into "1" and the other two possible answers are recoded into "0". The demographic characteristics that we use to predict treatment effects are shown in Figure 2. Because the variables we use for prediction can be considered categorical, we use the one-hot encoding approach in which each value of a variable (except for one) are represented as separate binary variables. We use these binary variables rather than the original features as our predictors.

Let Y be the outcome of interest and W the treatment group assignment, such that both Y and W have possible values in $0, 1$. We can define the average treatment effect (ATE) as

$$\mathbb{E}[Y = 1 \mid do(W = 1)] - \mathbb{E}[Y = 1 \mid do(W = 0)] \tag{1}$$

where $do(W = w)$ denotes *setting* as opposed to *observing* the value of the treatment condition [16, 17]. As we have argued above, a positive ATE does not guarantee that the treatment did not backfire for some populations. To find populations with different treatment effects, we consider the conditional average treatment effect (CATE), which is defined as

$$\begin{aligned} \tau =& \mathbb{E}[Y = 1 \mid do(W = 1), X = x] - \\ & \mathbb{E}[Y = 1 \mid do(W = 0), X = x] \end{aligned} \tag{2}$$

where X represents some vector of covariates, such as the demographic characteristics of a participant in a study. Figure 4 shows CATEs for different ethnic and age groups.

To estimate the CATEs, in our main results we use the T-learner approach [11]. This learner estimates two separate response functions: one for the outcome under the control condition and one for the outcome under the treatment condition. Formally, the response under control is defined as

$$\mu_0 = \mathbb{E}[Y = 1 \mid do(W = 0), X = x] \tag{3}$$

and the response under treatment is defined as

$$\mu_1 = \mathbb{E}[Y = 1 \mid do(W = 1), X = x] \tag{4}$$

The T-learner estimates μ_0 by predicting Y as a function of X using the control observations only; and it estimates μ_1 by predicting Y as a function of X using the treatment observations only. The difference between these estimates is then taken as the estimated CATE:

$$\hat{\tau} = \hat{\mu_1} - \hat{\mu_0} \tag{5}$$

The model we use to estimate the response functions is the gradient boosted tree algorithm in the scikit-learn machine learning library. [6, 18] We use default hyperparameters in all of our models.

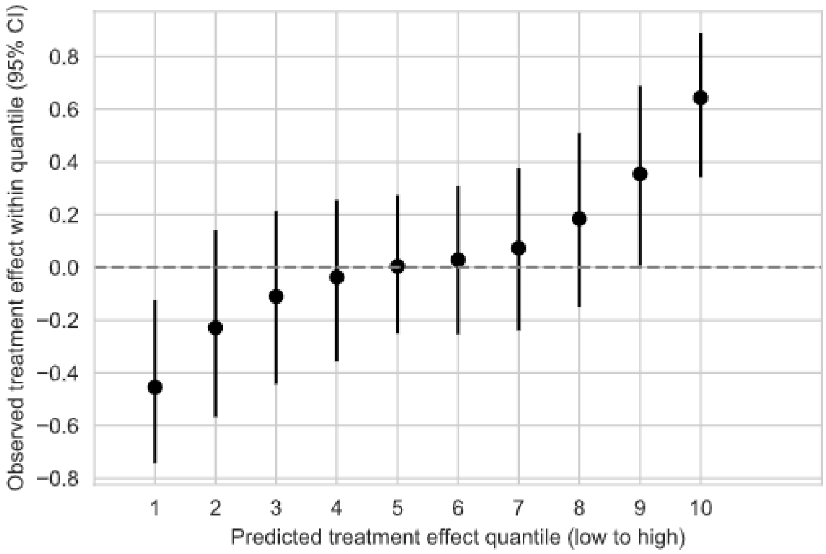

Figure 1: Observed treatment effects in testing samples sorted according to predicted treatment effects. For each quantile, the distribution of observed treatment effects is obtained by making predictions using 1000 bootstrapped datasets. The figure shows a clear correlation between predicted and observed treatment effects, indicating that our machine learning model is able to distinguish those individuals who are likely to respond to the advertisement campaign negatively from those who are likely to respond positively.

We evaluate the T-learner using a bootstrapping approach [5]. We first form 1000 new populations of size $N = 1600$ sampled (with replacement) from the original population and then divide each population into training and testing sets using a 80%/20% split. For each population, we train the T-learner on the training set and make predictions $\hat{\tau}$ on the testing set. The testing set is then divided into 10 quantiles according to the predicted $\hat{\tau}$. We use 10 quantiles to ensure a sufficient number of treatment and control observations within each group. We calculate the *observed* difference between those in the treatment group and those in the control group within each quantile and collect these observed differences across all 1000 bootstrap samples. We then calculate the treatment effects and empirical 95% confidence intervals for all 10 quantiles (shown in Figure 1). To calculate feature importances, we model $\hat{\tau}$ as a function of X in each of the 1000 bootstrap samples using a separate gradient boosting regressor. We then calculate the Gini feature importances [13] for each of these 1000 gradient boosting regressors and average the importance scores. Additionally, because the features are one-hot encoded, we average the importance scores across the different categories of each feature.

Finally, to ensure the robustness of our results, we repeat the entire analysis by using a different CATE learning method [14, 4]. We also conduct a follow up study to examine the relationship between climate beliefs and relevant demographic characteristics using ordinary least squares regression. The outcomes of these supplemental analyses, which are in line with the main results presented in the next section, are shown in the Appendix.

Results

The results of our analysis show clear population segments for whom the advertisements worked exceptionally well but also others where the ads clearly backfired (Figure 1). Indeed, the 10% of individuals with the highest predicted treatment effects increased their belief in climate change by an average of 64% whereas the 10% of individuals with the most negative treatment effects *reduced* their belief in climate change by 45%. This result shows that the exact same campaign can have either highly positive or highly negative effects depending on who you target.

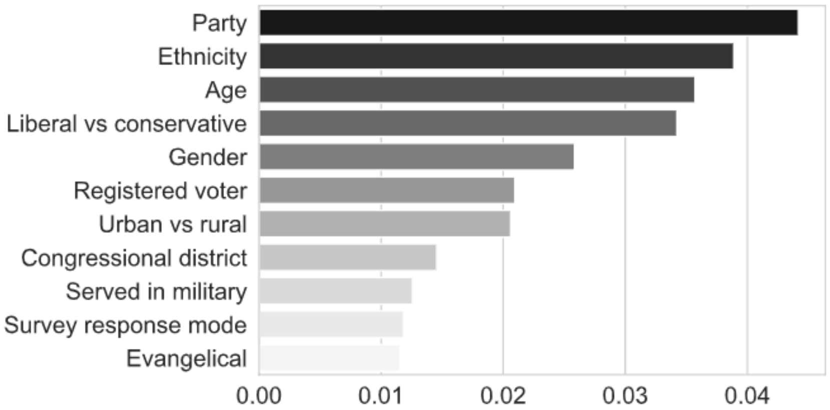

Figure 2: Feature importance scores for each of the 11 variables that we use to predict conditional average treatment effects. Party affiliation, ethnicity and age are the three variables most predictive of treatment effect heterogeneity.

We next examine the extent to which each of the 11 demographic characteristics predict differences in treatment effects by using using the variable importance based approach described above.This analysis shows that a number of characteristics predict the treatment effects in addition to party affiliation, including ethnicity, age, political ideology and sex (Figure 2).

Importance scores do not in and of themselves tell us how features are related with treatment effects. However, we can examine differences in characteristics between subgroups with the most positive and most negative predicted treatment effects. As reported by [7], our analysis finds clear differences in party affiliation, such that Republicans are more likely to respond positively to the ads than those with other party affiliations. However, when we look at the next two important characteristics, namely ethnicity and age, we find that younger, non-white populations responded more positively to the ads, whereas the ads were more likely to backfire in middle-aged and white populations.

The demographic differences that we observe by comparing the groups with the most negative and most positive predicted treatment effects are not always visible when we observe the entire population. This is because the differences tend to get diluted when we include the rest of the sample in our comparison. However, given the very pronounced differences in demographic characteristics that we find in our machine learning based analysis, we conduct a follow up investigation to examine the relationship between the treatment condition, ethnicity and age in the experimental population as a whole. For this analysis, we use a standard multiple regression approach. As proposed by our machine learning analysis, the population-level multiple regression model also indicates that age and ethnicity interact with the treatment effect. Figure 4 shows the difference between the treatment and control condition for different age groups and ethnicities. These results show that the differences observed by examining those with the most positive and most negative predicted responses are so strong that they are also visible at the population level. For details on the multiple regression results, see the Appendix.

Discussion

Behavioral scientists have long suggested that the effectiveness of climate change interventions should be evaluated with respect to specific demographic groups [25]. However, the methods researchers have traditionally used, such as examining main effects or conducting *ad hoc* moderation tests, often provide an incomplete view of the influence an intervention can have on different population segments. At the same time, causal machine learning methods have progressed rapidly and found promising applications in the tech industry [23] and personalized medicine [21]. Here, we show that causal machine learning can play a role in the mitigation of climate change, by improving the way in which communications about the subject are targeted. Specifically, our analysis shows that personalized ads referring to traditional Republican values worked well for non-white younger populations but backfired for white, middle-aged populations.

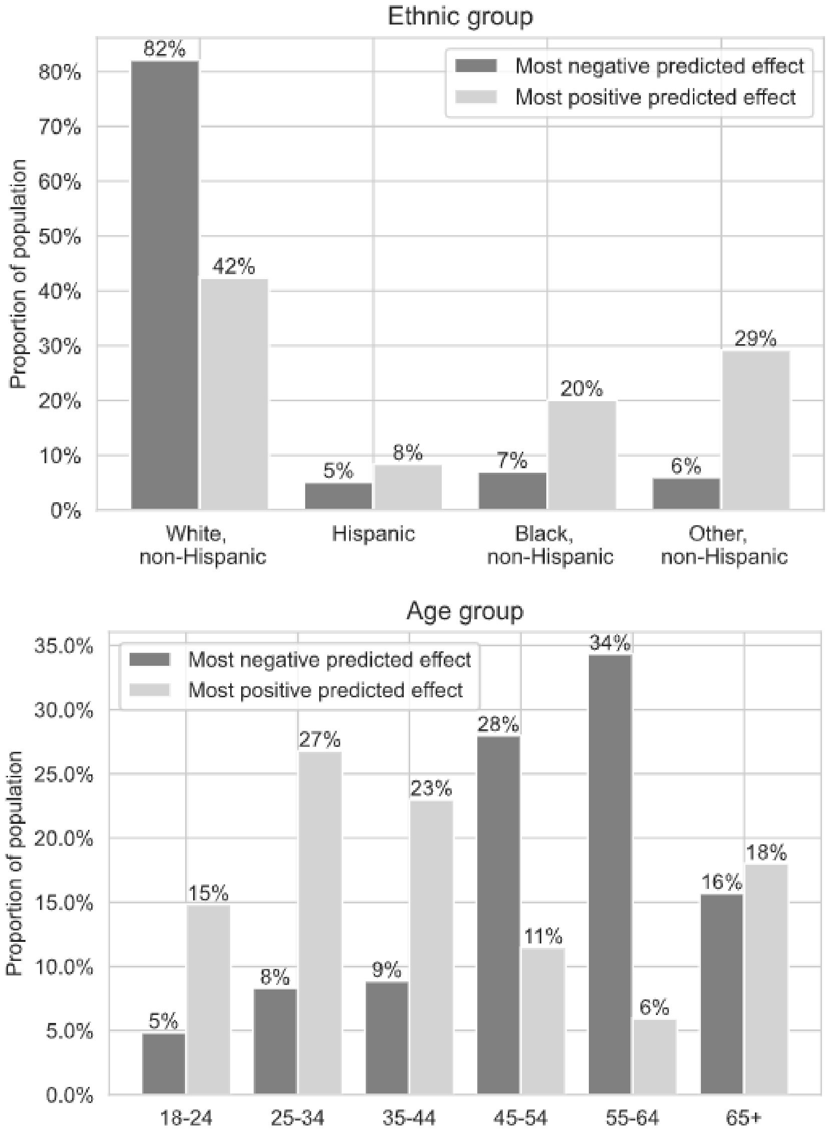

Figure 3: Ethnicity and age characteristics by predicted response to the advertising campaign. The comparison is between the 10% of the sample with the most negative predicted response and the 10% of the sample with the most positive predicted response. The proportions represent the averages in 1000 bootstrapped samples. The figure shows that those with the most negative predicted response tend to be white and middle-aged whereas those with the most positive predicted response tend to be non-white and younger.

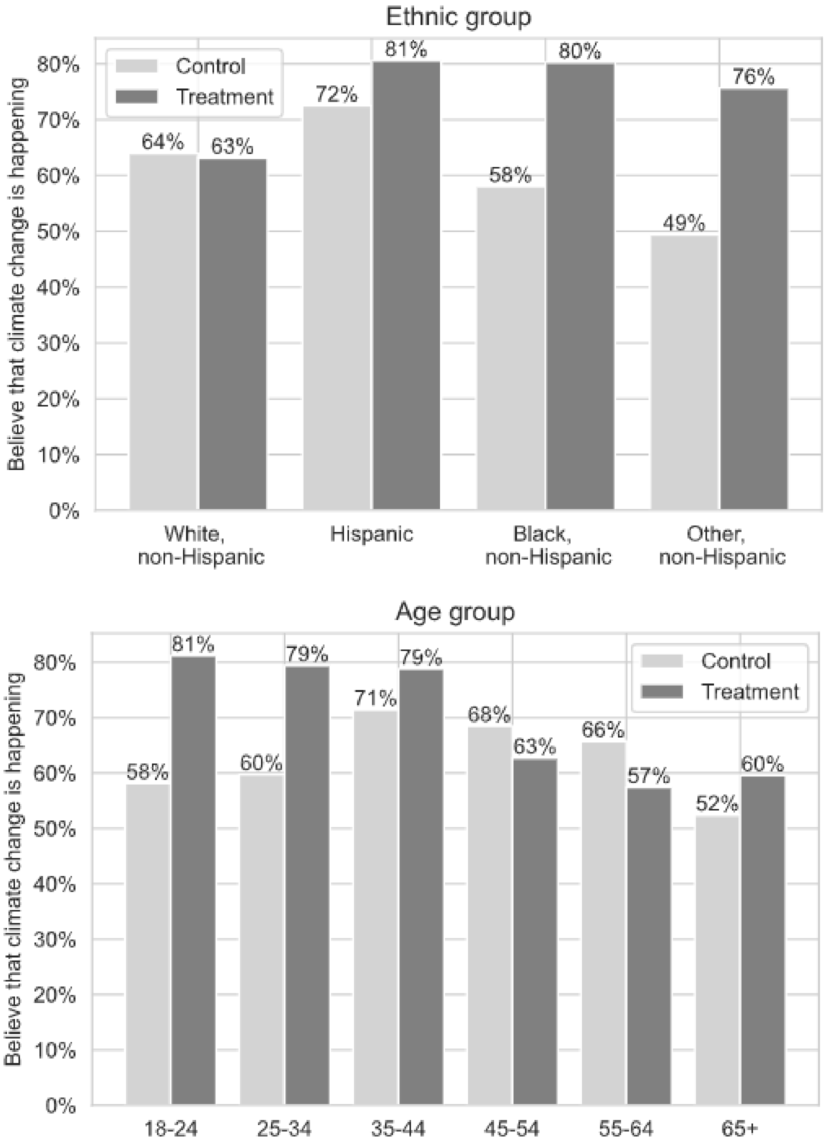

Figure 4: The relationship between the treatment effect and two demographic characteristics, ethnicity and age. The figures show that the relationships suggested by our machine learning model also hold if we examine the experimental population as a whole.

For policymakers, our results suggest a strategy where ads are first piloted in randomized experiments at a smaller scale. A targeting model can then be trained on the treatment effects observed in the pilot experiments, so that in the full campaign the advertiser can selectively target populations who are predicted to have the desired change in belief, and can avoid targeting populations for whom the campaign is predicted to backfire [1]. The analysis we present in this article generally agrees with [7] that personalized online advertising can play an important role in shifting views about climate change. However, our analysis also demonstrates that more modern methods, such as the causal machine learning approach we use in our analysis, are needed to have a more comprehensive understanding of the full influence of an intervention, and increase the likelihood that a intervention will succeed while reducing the chances that it will backfire.

Acknowledgments and Disclosure of Funding

We would like to thank Gill Pratt, Eric Krotkov, Brian Storey, Charlene Wu and Kent Lyons for comments. We would also like to thank anonymous reviewers of earlier versions for helpful suggestions. We have no conflicts of interest or additional sources of funding to disclose.

References

[1] Eva Ascarza. "Retention futility: Targeting high-risk customers might be ineffective". In: *Journal of Marketing Research* 55.1 (2018), pp. 80–98.

[2] Susan Athey and Guido Imbens. "Recursive partitioning for heterogeneous causal effects". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 113.27 (July 2016), pp. 7353–7360.

[3] Christopher J Bryan, Elizabeth Tipton, and David S Yeager. "Behavioural science is unlikely to change the world without a heterogeneity revolution". In: *Nature human behaviour* 5.8 (2021), pp. 980–989.

[4] Huigang Chen et al. "Causalml: Python package for causal machine learning". In: *arXiv preprint arXiv:2002.11631* (2020).

[5] B Efron. "Bootstrap Methods: Another Look at the Jackknife". en. In: *aos* 7.1 (Jan. 1979), pp. 1–26.

[6] Jerome H Friedman. "Greedy function approximation: A gradient boosting machine". en. In: *aos* 29.5 (Oct. 2001), pp. 1189–1232.

[7] Matthew H Goldberg et al. "Shifting Republican views on climate change through targeted advertising". en. In: *Nat. Clim. Chang.* (June 2021), pp. 1–5.

[8] Robin Gomila. "Logistic or linear? Estimating causal effects of experimental treatments on binary outcomes using regression analysis". en. In: *J. Exp. Psychol. Gen.* 150.4 (Apr. 2021), pp. 700–709.

[9] Pierre Gutierrez and Jean-Yves Gérardy. "Causal Inference and Uplift Modelling: A Review of the Literature". In: *Proceedings of The 3rd International Conference on Predictive Applications and APIs*. Ed. by Claire Hardgrove et al. Vol. 67. Proceedings of Machine Learning Research. Microsoft NERD, Boston, USA: PMLR, 2017, pp. 1–13.

[10] Behram Hansotia and Brad Rukstales. "Incremental value modeling". en. In: *J. Interact. Mark.* 16.3 (Jan. 2002), pp. 35–46.

[11] Sören R Künzel et al. "Metalearners for estimating heterogeneous treatment effects using machine learning". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 116.10 (Mar. 2019), pp. 4156–4165.

[12] Ang Li and Judea Pearl. "Unit selection based on counterfactual logic". In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*. Macao, China: International Joint Conferences on Artificial Intelligence Organization, Aug. 2019.

[13] Bjoern H Menze et al. "A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data". en. In: *BMC Bioinformatics* 10 (July 2009), p. 213.

[14] Xinkun Nie and Stefan Wager. "Quasi-oracle estimation of heterogeneous treatment effects". In: *Biometrika* 108.2 (2021), pp. 299–319.

[15] Brendan Nyhan et al. "Effective messages in vaccine promotion: a randomized trial". en. In: *Pediatrics* 133.4 (Apr. 2014), e835–42.

[16] Judea Pearl. *Causality*. Cambridge University Press, Sept. 2009.

[17] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

[18] Fabian Pedregosa et al. "Scikit-learn: Machine learning in Python". In: *the Journal of machine Learning research* 12 (2011), pp. 2825–2830.

[19] Nicholas J Radcliffe and Patrick D Surry. "Real-world uplift modelling with significance-based uplift trees". In: *White Paper TR-2011-1, Stochastic Solutions* (2011), pp. 1–33.

[20] Piotr Rzepakowski and Szymon Jaroszewicz. “Decision trees for uplift modeling with single and multiple treatments”. In: *Knowledge and Information Systems* 32.2 (2012), pp. 303–327.

[21] Pedro Sanchez et al. *Causal Machine Learning for Healthcare and Precision Medicine*. 2022. DOI: `10.48550/ARXIV.2205.11402`. URL: `https://arxiv.org/abs/2205.11402`.

[22] Briony Swire-Thompson, Joseph DeGutis, and David Lazer. “Searching for the Backfire Effect: Measurement and Design Considerations”. en. In: *J. Appl. Res. Mem. Cogn.* 9.3 (Sept. 2020), pp. 286–299.

[23] Vasilis Syrgkanis et al. “Causal Inference and Machine Learning in Practice with EconML and CausalML: Industrial Use Cases at Microsoft, TripAdvisor, Uber”. In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery Data Mining*. KDD ’21. Virtual Event, Singapore: Association for Computing Machinery, 2021, pp. 4072–4073. ISBN: 9781450383325. DOI: `10.1145/3447548.3470792`. URL: `https://doi.org/10.1145/3447548.3470792`.

[24] Stefan Wager and Susan Athey. “Estimation and Inference of Heterogeneous Treatment Effects using Random Forests”. In: *J. Am. Stat. Assoc.* 113.523 (July 2018), pp. 1228–1242.

[25] Elke U Weber and Paul C Stern. “Public understanding of climate change in the United States.” In: *American Psychologist* 66.4 (2011), p. 315.

[26] Yan Zhao, Xiao Fang, and David Simchi-Levi. “Uplift Modeling with Multiple Treatments and General Response Types”. In: *Proceedings of the 2017 SIAM International Conference on Data Mining (SDM)*. Proceedings. Society for Industrial and Applied Mathematics, June 2017, pp. 588–596.

[27] Zhenyu Zhao and Totte Harinen. “Uplift Modeling for Multiple Treatments with Cost Optimization”. In: *2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. ieeexplore.ieee.org, Oct. 2019, pp. 422–431.

A Appendix

The tables below show the ordinary least squares regression [8] results on the relationships between the treatment condition, age and ethnicity. Given the large differences in age and ethnicity between those with the most positive and most negative treatment effects, we conducted OLS regression as a follow-up study to see if the differences would be visible at the level of the population as a whole. The results how that the differences were indeed visible when looking at the sample as a whole.

	coef	**std err**	**t**	**P> |t|**
Intercept	0.6397	0.019	32.810	0.000
Ethnicity[Hispanic]	0.0853	0.078	1.098	0.272
Ethnicity[Black]	-0.0595	0.056	-1.057	0.291
Ethnicity[Other]	-0.1456	0.055	-2.642	0.008
Condition	-0.0083	0.027	-0.303	0.762
Condition:Ethnicity[Hispanic]	0.0889	0.113	0.790	0.430
Condition:Ethnicity[Black]	0.2305	0.080	2.898	0.004
Condition:Ethnicity[Other]	0.2706	0.079	3.408	0.001

R-squared:	0.018	**F-statistic:**	4.152
Df Residuals:	1592		

Table 1: OLS regression results for the relationship between ethnicity and the treatment condition. The dependent variable is belief in climate change. The contrast category for Ethnicity is is “White”.

	coef	**std err**	**t**	**P> |t|**
Intercept	0.5811	0.055	10.592	0.000
Age[25-34]	0.0163	0.077	0.212	0.832
Age[35-44]	0.1332	0.071	1.884	0.060
Age[45-54]	0.1035	0.067	1.542	0.123
Age[55-64]	0.0762	0.065	1.168	0.243
Age[65+]	-0.0573	0.064	-0.898	0.369
Condition	0.2307	0.075	3.074	0.002
Condition:Age[25-34]	-0.0340	0.103	-0.328	0.743
Condition:Age[35-44]	-0.1571	0.099	-1.581	0.114
Condition:Age[45-54]	-0.2884	0.094	-3.076	0.002
Condition:Age[55-64]	-0.3138	0.091	-3.441	0.001
Condition:Age[65+]	-0.1589	0.088	-1.814	0.070

R-squared:	0.034	**F-statistic:**	5.054
Df Residuals:	1588		

Table 2: OLS regression results for the interaction between age and the treatment condition. The dependent variable is belief in climate change. The contrast category for Age is "18-24".

Figure 5 shows the results of following the same analysis procedure as described in the Methods section, except for the CATE learning algorithm, which is here the R-learner [14]. This is to ensure that the results remain robust irrespective of the specific way in which we learn treatment effect heterogeneity. As the base-learners in the R-learner, we used the same gradient boosting implementation and parameters as we did in the T-learner reported in the main results. Additionally, because the R-learner requires the propensity score $P(W = 1 \mid X = x)$ as an input, we used L2-penalised logistic regression with 4-fold crossvalidation to learn that paramater. As can be seen, the results of this analysis resemble very closely those obtained using the T-learner.

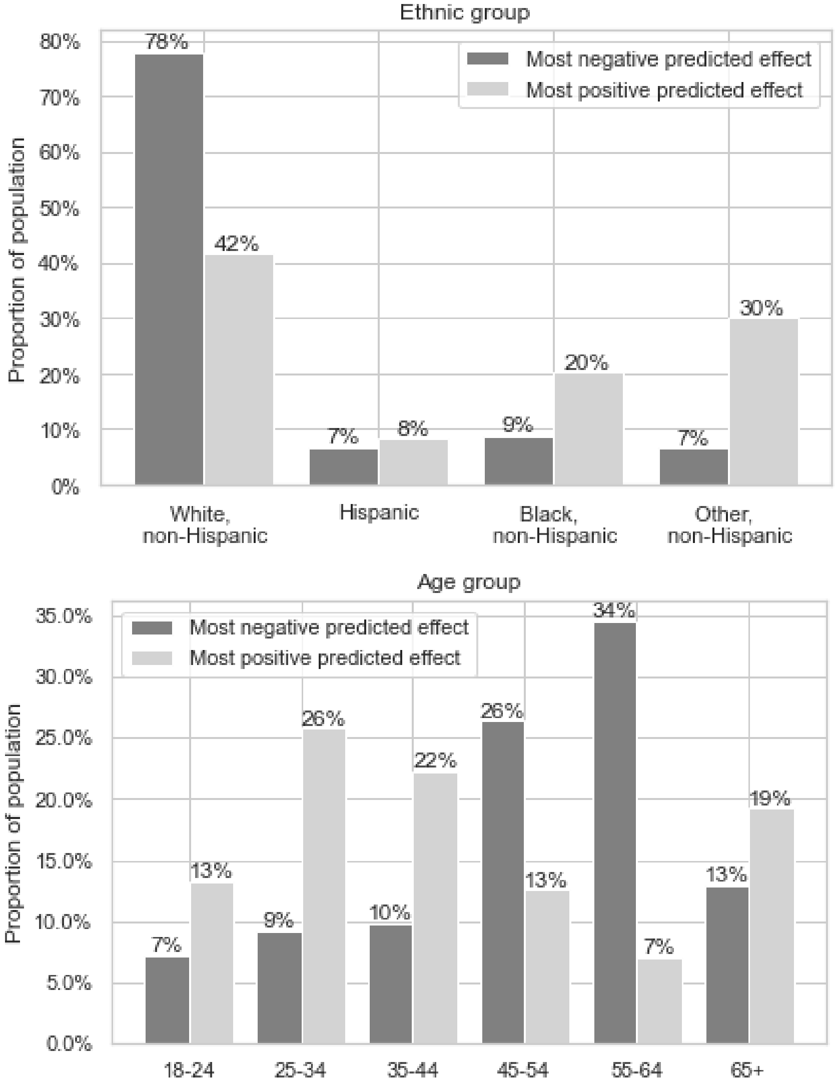

Figure 5: Ethnicity and age characteristics by predicted response to the advertising campaign, using the R-learner. Apart from the CATE learning algorithm, the analysis steps were identical to those used to obtain the main results. As can be seen by comparing the demographic distributions shown here to those shown in Figure 3, the results obtained using the R-learner are very similar to those obained using the T-learner.