Does Micro-Targeting Work? Evidence from an Experiment during the 2020 United States Presidential Election

Musashi Harukawa, University of Oxford

Abstract

Micro-targeted political campaigning continues to be a salient topic both in scholarship and public debate. Despite a diverse literature warning about the dangers this technology poses to elections, democracy, and civic society, extant political science research indicates it may have no effect. This paper leverages a novel design incorporating both machine learning and causal inference to estimate the effect of micro-targeted political advertising. Among unaligned voters who had not cast their vote at the time of the survey, targeting anti-Biden advertisements run by the Trump campaign increased reported Biden dislike by 8.7 percentage points, and decreased intent to vote for Biden by 7.1 percentage points compared to randomly assigned anti-Biden advertisements. The magnitude of these estimates indicates that micro-targeting may be able to affect the outcome of elections.

Introduction

Does micro-targeting work? After the Brexit referendum and election of Donald Trump in 2016, observers raised questions about the role of sophisticated digital campaign strategies employed by *Cambridge Analytica* and others (Simon 2019). Watchdogs, journalists and scholars identified a raft of issues stemming from these technologies: threats to privacy from the systematic harvesting of individual data by corporations (Zuboff 2015; Wachter 2017), threats to civil discourse from the creation of informational "filter bubbles" (Burkell and Regan 2020), to threats to autonomy and democracy from mass manipulation (Sunstein 2015). Meanwhile, attempts to curb this practice vary, with the practice alive and well in the United States (Wood and Ravel 2017; Dobber, Ó Fathaigh, and Zuiderveen Borgesius 2019).

This vibrant normative, legal and ethical debate widely assumes that these new technologies are effective; "micro-targeting of voters can pay very handsome electoral dividends for a relatively modest investment" (Krotoszynski 2020). Recent political science literature, however, casts doubt on this assertion (Nickerson and Rogers 2020). A large panel study by Coppock, Hill, and Vavreck (2020) testing 49 advertisements from the 2016 United States presidential election campaigns finds not only that average persuasive effects are small, but that a lack of heterogeneous effects leaves little room for targeting to work. On the other hand, psychological models of persuasion tell us that campaigns that know characteristics of their receivers can increase the effectiveness by catering to those features (Cialdini 2007; Madsen and Pilditch 2018). Efforts to reconcile these results are hampered by the challenges of studying campaigns and their activities online (Baldwin-Philippi 2017; Edelson et al. 2019).

Approaching the puzzle from a machine learning and causal inference angle, this paper presents a novel two-stage survey experiment designed to causally estimate the effect of micro-targeted political campaigning. Using five real anti-Biden advertisements fielded by the Donald Trump campaign in the final month of the 2020 United States presidential election, the design leverages the randomized assignment in the first stage to train an on-line targeting algorithm, which is then used to optimally allocate advertisements on the basis of respondent characteristics in the second stage. Assuming the participants in the two stages are comparable, we can obtain a consistent estimate of the effect of micro-targeting by comparing anti-Biden attitudes and other outcomes between the two stages.

The results show that the optimized allocation of advertisements leads to a significant boost in anti-Biden sentiment and intent to not vote for Biden among undecided voters who had not already voted at the time of the survey. Although the overall ATE of targeting was found to be small, among voters who identified with neither party and had not voted at the time of the survey, the effect size is a substantial. In this group, targeting increased the proportion of who disliked Biden by 8.7 percentage points, and decreased the proportion intending to vote for Biden by 7.1 percentage points.

The magnitude of these estimates implies that micro-targeting could change the result of very close elections with a sufficient number of unaligned voters, such as Arizona and North Carolina in the 2020 United States presidential election. This paper concludes by discussing the normative and substantive implications of these findings, and outlines future research that can provide a template for the effective regulation of this technology.

Context

Tailoring and *targeting* are two key concepts that I use extensively in this paper and whose meaning vary within the literature. *Tailoring* a message is constructing it in such a way that it is designed to appeal to a specific audience. *Targeting* refers to delivering the message in such a way that only the intended audience sees it (Burkell and Regan 2020; Fowler et al. 2021). Although the prefix “micro-” is not explicitly defined in the literature, I use "micro-targeting" to refer to targeting being done on the basis of individual-level data, usually obtained via social media or a data broker. This article focuses on the effect of the advertisement allocation strategies of data-driven political campaigns, or *political micro-targeting*.

Scholars have been interested in the use of targeted messaging in political campaigning since at least 2005 (Kang 2005), but political micro-targeting came under renewed public and scholarly attention following major political events in 2016 and claims that the political data analytics firm CA had a role in these outcomes (Simon 2019).

In particular, concern has been directed towards the ability of political campaigns to specify with a high degree of granularity the traits of the recipient that should receive a given advertisement on online platforms such as Facebook (Zuboff 2015; Baldwin-Philippi 2017). Combined with the extensive data on voters available to corporations and campaigns (Wachter 2017), these tools are one of the myriad ways that political campaigns develop and deliver sophisticated messaging that combines data on recipient traits with the ability to selectively show narrowly tailored messages to those it is designed to affect (Fowler et al. 2021).

Legal and ethical scholars have highlighted the potential harms of this technology, from the creation of informational filter bubbles (Pariser 2011) that hamper civil discourse (Burkell and Regan 2020), to the incentivization of data harvesting by public officials (Krotoszynski 2020). A number of scholars identify systemic threats to democracy, including the disenfranchising nature of manipulative advertising (Sunstein 2015; Burkell and Regan 2019), and difficulty of countering these changes through legislative means (Wood and Ravel 2017; Hindman 2018).

These arguments by and large assume that micro-targeting *works*. Nevertheless, there are strong reasons to be skeptical of the claims of campaigners and their ability to manipulate the masses (Nickerson and Rogers 2020). A decade of political science field experiments looks at the effect of political advertising, whether on television and radio (Gerber et al. 2011; Mitchell 2012) or online (Broockman and Green 2014; Edelson et al. 2019), finding by and large that even where advertisements have significant effects, these effects rapidly decay and are quickly replaced by new information.

A recent large field experiment ($N = 34000$, Coppock, Hill, and Vavreck 2020) finds that the effect of advertisements is small and largely homogeneous. The authors conclude: "... expensive efforts to target or tailor advertisements to specific audiences require careful consideration. The evidence... shows that the effectiveness of advertisements does not vary greatly from person to person or from advertisement to advertisement" (Coppock, Hill, and Vavreck 2020, 6). Correspondingly, targeting the delivery of advertisements should be unlikely to yield any gains in effectiveness. It is difficult to square these null findings with claims of using micro-targeting for mass manipulation.

Meanwhile, studies from the field of psychology on *psychometric profiling*, the technique employed by CA, find evidence that gains should be possible. Psychometric profiling is the technique of building psychological "profiles" of voters, which are then used to develop advertisements tailored to different types of profiles. Underlying this technique are psychological models of persuasion (e.g. Petty and Cacioppo 1986; Cialdini 2007), which emphasize that whether a message is processed *emotively* or *cognitively*[1] depends on an interaction between message content, receiver characteristics and receiver social context. Knowledge of receiver characteristics allows a campaign to tailor an advertisement to increase the likelihood that the message will be processed in a manner that results in opinion or attitude change.

Building on these models, Madsen and Pilditch (2018) use agent-based modelling to demonstrate that a campaign with more information about voter characteristics should be more successful in the election. The results of this simulation are bolstered by experimental evidence; Zarouali et al. (2020) use a personality profiling algorithm to label participants as introverts or extroverts, and then show them a personality-congruent advertisement based on this label, finding that correctly matched advertisements have a stronger effect.

Whereas these studies assess the effectiveness of a particular strategy for tailoring and targeting in conjunction, it remains difficult to determine the extent to which this strategy is representative of contemporary campaigning practices. The relevant actors in this space have little incentive to share truthful and complete information regarding their activities; campaigns and consultancies will likely claim that their proprietary techniques work (Simon 2019; Hindman 2018), and attempting to collect data on the activities of these advertisers is hampered by the platforms themselves (Edelson et al. 2019)[2].

In contrast to the above studies, I adopt an inductive approach to modelling micro-targeting (Grimmer, Roberts, and Stewart 2021); instead of assuming a particular approach to targeted campaigning, I fit multiple models and employ the one that best predicts my data. This study aims to focus solely on whether targeting can be consequential by testing the key mechanism underlying micro-targeted campaigning. By determining whether it is possible for a campaign to improve their performance by optimizing allocation of advertisements on the basis of receiver traits, this experiment shows whether there are possible gains from micro-targeting. If gains are possible, then the consequences of this technology for democratic politics, from undermining elections (Wood and Ravel 2017; Burkell and Regan 2020) to the incentives for the abuse of data by elected officials (Krotoszynski 2020), are considerable.

[1] In the language of the *Elaboration Likelihood Model* (Petty and Cacioppo 1986), the modes of processing information range on a continuum from central (critical, rational) to peripheral (heuristic, emotive).

[2] See also, for instance, the efforts of Facebook to prevent researchers at NYU and the newsroom ProPublica from collecting data on political advertisements being shown to users.

Experimental Design

The choice of the 2020 United States presidential election as the setting for this experiment was motivated by several factors. First and foremost, it was my aim to use real advertisements that had been tailored as part of a targeted advertising campaign. Enormous campaign spending and the lack of regulation of targeted political advertising (Baldwin-Philippi 2017) makes the United States a better candidate than European democracies for finding such advertisements (Dobber, Ó Fathaigh, and Zuiderveen Borgesius 2019). Similarly, the higher stakes in the presidential election increased the likelihood of granular targeting campaigns while simultaneously providing the entire national audience as the relevant population to sample from.

I begin by outlining the requirements of an experimental design for estimating the average effect of targeting. Given N respondents and a advertisements, we split the respondents into a treatment and control group and show each of them exactly one advertisement. The control group receives an untargeted advertisement; in other words, the assignment is independent of respondent characteristics. The control group receives a targeted advertisement; assignment is made to maximize predicted effect as a function of respondent characteristics. This optimal assignment is made possible by using the data from the control group to train a predictive model, and then using the pre-treatment responses to generate allocations in real time. Provided that assignment to treatment/control is random, the difference in average outcome between the treatment and control group is an unbiased estimate of the average effect of targeting.

Because the design requires the data from the control group in order to train the targeting algorithm, I refer to these groups in their sequential order, stage 1 (control) and stage 2 (treatment). This naming is meant in part to emphasize that assignment to these groups was done sequentially, inducing possible time-of-day effects.[3]

Each of the respondents $i \in N$ was allocated one of five advertisements $D_{i,a}$, $a \in [1, 5]$. The five advertisements, detailed in Table 1: Advertisements and Descriptions, were selected from the set of all anti-Biden advertisements run by the Donald Trump campaign on Facebook and YouTube. The criteria for selection are: all advertisements are clearly tailored to different audiences, but no single advertisement is likely to out-perform all others. Advertisements focusing on Joseph Biden were chosen in favor of Donald Trump because two weeks prior to the election it was uncertain would Trump would drop out due to COVID-19. Although Coppock, Hill, and Vavreck (2020) find no evidence of asymmetric effects for attack versus promotional advertisements, attack advertisements were chosen because there is an extensive literature focusing on the normative issues surrounding negative advertising (Ansolabehere and Iyengar 1995).

[3] In order to mitigate time-of-day effects, I conducted the survey in as small a timespan as possible. The results of balance checks are discussed in the results section, and other robustness checks are detailed in *Appendix III: Implementation Notes.*

Table 1: Advertisements and Descriptions

Title	Description
They Mock Us	Clinton and Biden are mocking you (In-Group)
Why did Biden let him do it?	Hunter Biden's ostensible corruption
Biden will come for your guns	Second Amendment; Biden will steal guns
Insult	Biden: Black Trump supporters not Black
Real Leadership	Obama/Biden caused wars, neglected veterans

In the first stage, respondents first answered demographic and political questions regarding age, gender, race, income group, state, interest in news, whether they thought the country was on the right track for the past four years, a seven-point partisan identification scale, and a five-point liberal-conservative ideological self-identification scale. The format and wording of these items was adopted from the American National Election Study.

Respondents were then given a prompt stating that they would be shown a short advertisement, and then served one of the five advertisements detailed above at random. Randomization was done using permuted block randomization over a discrete uniform distribution, i.e. $a \sim \mathcal{U}\{1, 5\}$.

After viewing the advertisement, respondents were asked three post-treatment items (plus a manipulation check). These included their favorability rating of Biden and Trump on one to five scale, and their voting intention. Given that the survey was run very close to the actual election and there were high rates of early voting, respondents were given the option to state whom they had already voted for.

The data gathered in the first stage was then used to train a predictive model to learn Biden favorability, Y_i, as a function of the pre-treatment covariates and advertisement shown, $f(X_i, D_a)$. Thus given any profile of pre-treatment covariates, the predicted outcome under each of the five advertisements would be calculated $\{\hat{Y}(D_1), \hat{Y}(D_2), ...\hat{Y}(D_5)\}$, and the advertisement that minimized[4] Biden favorability, $D^*(X_i)$, would be allocated:

$$D^*(X_i) := argmin_a \; f(X_i, D_{i,a})$$

Five candidate models were fitted during the experiment to learn this relationship. The first three, all algorithms over decision trees, were chosen primarily for their ability to learn highly conditional response

[4] Note that if this advertisements were promotional, then $D^*(X_i)$ would be the value that *maximizes* predicted favorability.

surfaces: Random Forest (RF), AdaBoost and Gradient Boosted Decision Trees (GBDT). Additionally, a Multi-Layer Perceptron Regressor (MLPR) and Support Vector Machine (SVM) were included for comparison, with the latter hedging on the possibility that the response surface was in fact best approximated with a linear kernel. All five of these models are standard, "off-the-shelf" algorithms included in the widely used `scikit-learn` library (Pedregosa et al. 2011), in order to show that campaigns with a relatively low level of data science sophistication could easily implement similar approaches.

The five fitted models were compared on root mean squared error (RMSE), maximum error and prediction time[5], in order to find the model best able to give quick and accurate predictions of outcome under alternative advertisement allocation. RF and AdaBoost outperformed the latter three on all of these metrics. Between RF and AdaBoost, RF performed weakly better and was less likely to predict ties, and was thus prefrable to AdaBoost for predicting optimal advertisements[6]. The pre-fitted RF model was then uploaded to the web server.

In the second stage, respondents first answered the same pre-treatment questions. These answers were sent to a server-side `Python` kernel, which given the pre-treatment covariates X_i, used the pre-fitted RF model to allocate the optimal advertisement $D^*(X_i)$. The respondent then watched the advertisement and answered the same three post-treatment items. The first and second stage were therefore indistinguishable from the perspective of the respondent, as they were unaware which advertisements they were not shown and how the advertisement was allocated to them until an end-of-survey debrief.

Provided that there are no systematic differences between the first and second stage, we can compare the outcome between randomly assigned stage 1, $\mathbb{E}_a[\mathbb{E}_i[Y_i(D_{i,a})]]$, and optimally assigned stage 2, $\mathbb{E}_i[Y_i(D^*(X_i))]$[7], as an estimator of the average effect of targeting. As it is, we must account for possible time-of-day effects for this estimator to be unconfounded.

$$ATE = \mathbb{E}_i[Y_i(D^*(X_i))] - \mathbb{E}_a[\mathbb{E}_i[Y_i(D_{i,a})]]$$

Following a power analysis based on a simulated run of the experiment using the replication data for Coppock, Hill, and Vavreck (2020), the total N was set at 2400,[8] with 1500 respondents allocated to the control group

[5] Prediction time was important because the allocation engine could only handle single tasks, and the long queues during traffic bursts would have created uncontrolled variation in the respondent experience of the survey.

[6] In the case of ties, an optimum was randomly chosen.

[7] *Note on Notation:* In the first stage (right-hand term of equation), we average the value of the outcome over all individuals ($\mathbb{E}_i$) and an equal probability of seeing any particular advertisement ($\mathbb{E}_a$). In the second stage (left-hand term of equation), the advertisement allocation is not random, but a function of pre-determined respondent traits ($D^*(X_i)$), and therefore we only average over individuals.

[8] For further details of the simulation and power analysis, see Figure 3 in *Appendix I: Theoretical Notes.*

and 900 respondents allocated to the treatment group. All participants are United States citizens, resident in the United States, of voting age. Because the design requires a sizable sample to undertake the experiment in a short window of time, the experiment was conducted online rather than in-person or over the phone. Respondents were recruited via the survey provider Prolific and redirected to a custom-built website at `https://survey.polinfo.org`.

Hypotheses

This design is used to test the effect of targeting on three outcomes that a campaign is likely to want to influence. The first, *favorability*, is the degree to which a candidate is liked or disliked. The second, *voting preference* is which candidate they intend to vote for. The third, *turnout intention*, is whether the individual intends to vote at all. For respondent i, Biden favorability is denoted $y_i \in \langle 1, 2, 3, 4, 5 \rangle$, intent to vote for Biden is denoted v_i, and intent to vote at all is denoted u_i.

If micro-targeting is effective, then a campaign that optimally allocates advertisements on the basis of individual traits should be more successful in achieving its aims. Given that the advertisements being used are anti-Biden advertisements run by the Trump campaign, we get the following testable hypotheses:

- `Hypothesis 1` (*Micro-targeting Affects Favorability*): $\mathbb{E}_i[y_i(D^*)] < \mathbb{E}_a[\mathbb{E}_i[y_i(D_{i,a})]]$
- `Hypothesis 2` (*Micro-targeting Affects Voting Preference*): $\mathbb{E}_i[v_i(D^*)] < \mathbb{E}_a[\mathbb{E}_i[v_i(D_{i,a})]]$
- `Hypothesis 3` (*Micro-targeting Affects Turnout*): $\mathbb{E}_i[u_i(D^*)] < \mathbb{E}_a[\mathbb{E}_i[u_i(D_{i,a})]]$

Note that the latter two outcomes are only measurable for individuals who have not already voted at the time of the survey. Therefore the first hypothesis is tested for both the full sample and subset of voters who had not voted at the time of the survey, and the latter two hypotheses are only tested conditional on having not voted at the time of the survey.

I also look at the above three hypotheses conditioning on the effect of partisanship. There are a multitude of reasons to suspect that the treatment effect will be heterogeneous on partisanship. For one, an extensive psychology literature demonstrates that individuals are less likely to adopt attitudes that run contrary to their stable beliefs or unchangeable past actions (see e.g. Cialdini 2007). Moreover, strongly partisan respondents are likely to have a baseline level of Biden preference that is at either extreme of the 1 to 5 scale, meaning that even where the treatment does have an effect, it cannot be measured by the question format employed. Finally, campaigns themselves will typically segment and distinguish between supporters, opponents and on-the-fence voters, and employ their greatest persuasive effort trying to "swing" the middle. Thus we are *a priori* interested in the conditional effect of targeting on these different groups.

Results

The survey ran on 28 October 2020 between 5pm and midnight UTC. After filtering for irregularities[9], the experiment provided a total of 2261 valid responses, with 1416 in the control group and 845 in the treatment group. The results are presented in three parts. The first part examines the predictive model used to allocate advertisements. The second part provides checks for sample balance between the first and second stage. The third part evaluates the hypotheses listed in the previous section.

Predictive Model and Allocations

The four panels of Figure 1 explore different aspects of the predictive model used to allocate advertisements in stage 2. Panels 1 and 2 concern the predictive "logic" of the model, whereas panels 3 and 4 address the allocations themselves.

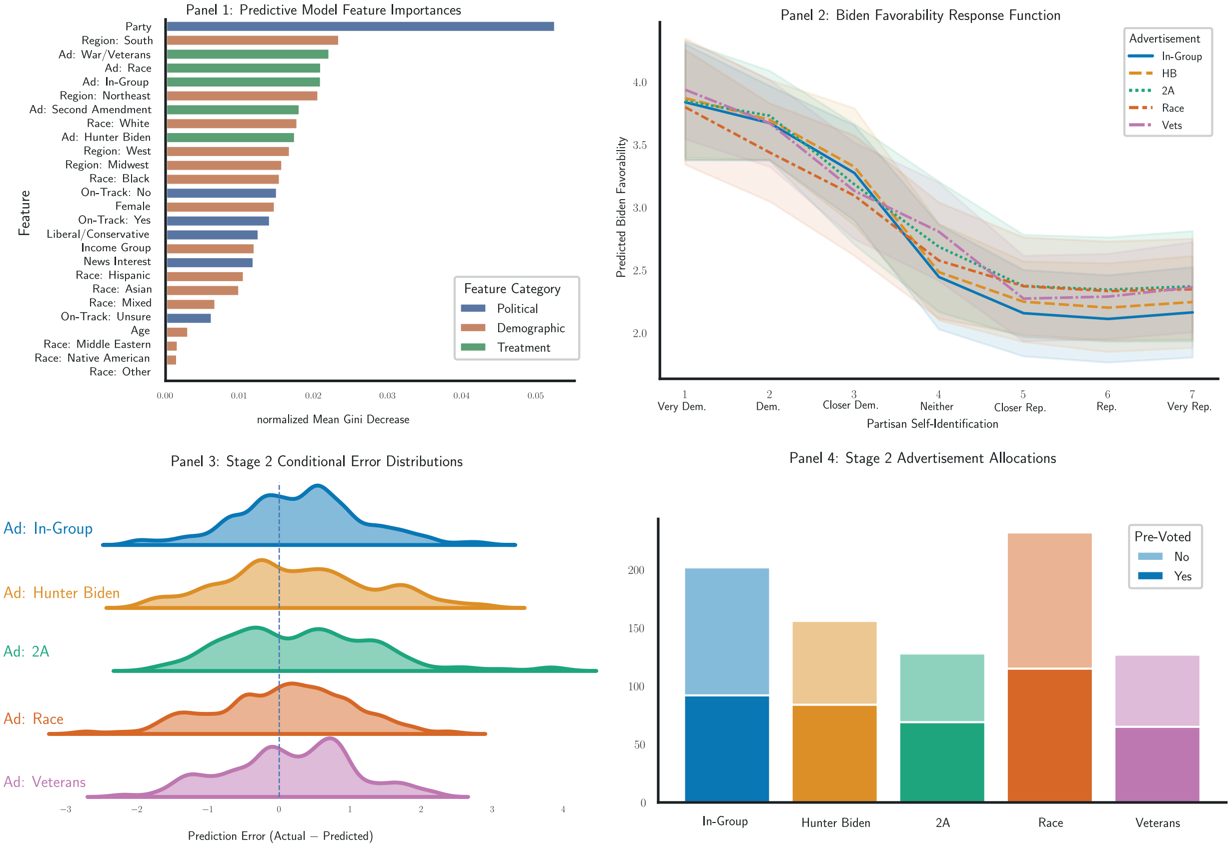

Figure 1: *Panel 1: feature importances for predictive model used to allocate optimal advertisements; Panel 2: average Biden favorability response curve as function of partisanship over stage 2 respondents; Panel 3: per-advertisement conditional prediction error distributions; Panel 4: stage 2 advertisement allocation counts.*

[9] Irregularities consisted of incomplete surveys or failed attention checks. For complete descriptions of manipulation checks and balance checks, see Appendix III.

A question of substantive interest to both researchers and campaigners alike is why the predictive algorithm assigned a particular advertisement and not another. Unlike ordinary least squares (OLS), the RF model employed does not have regression coefficients to facilitate interpretation of the predicted outcome as the sum of linear components. Nevertheless, decision-tree based algorithms do provide metrics that allow substantive interpretation (Montgomery and Olivella 2018).

Panel 1 shows the per-feature normalized Mean Gini Decrease[10] (nMGD) statistic from the RF model used to assign allocations in the second stage. nMGD tells us the degree to which partitions on a feature produce homogeneous sub-spaces at each node of the individual decision trees. In other words, this statistic tells us the extent to which a feature produces systematic and separable regions of the label space.

Features with high nMGD therefore provide much of the predictive power for RF models. In panel 1, it can be seen that partisan self-identification has the highest feature importance, followed by the respondent being from the South. What is notable, however, is that all of the advertisements have high feature importances. Substantively, this means that the algorithm "learned" more about a respondent's Biden favorability from which advertisement they were shown than their gender, income group, or race (with the exception of *Race: White* versus *Ad: Hunter Biden*).

Although feature importance does not correspond to any causal estimand, the randomized advertisement assignment in this first stage leaves no other explanation for high feature importance other than the heterogeneous effects between advertisements.[11]

It can be seen that at levels of partisanship ranging from *1: Very Democrat* to *3: Neither, but Closer to Democrat*, the RF model predicts that the "Race" advertisement (titled "Insult" in Table 1) gives the lowest predicted Biden favorability, and is therefore the most effective. At partisanship *4: Neither* and above, the most effective advertisement is predicted to be the In-Group advertisement (titled "They Mock Us" in Table 1), although at 4 the difference between the In-Group advertisement and Hunter Biden is negligible.

The shaded areas around each curve show the standard deviation of the response curve; it can be seen that the range covered by each of the curves is mostly overlapping. This reflects that the individual response functions vary greatly by other covariates, and that there is considerable heterogeneity in individual response to the advertisements.

Panel 3 shows the conditional error distributions of the predictions made by the model for the stage 2

[10] As MGD is biased towards features with high cardinality, I normalize the feature importances by their cardinality.

[11] A causal interpretation for the Stage 1 results and RF model is discussed more extensively in *Appendix I: Theoretical Notes*. Panel 2 shows the mean response curve as a function of partisanship for all stage 2 respondents. This was calculated by using the fitted RF model to predict Biden favorability for each of the stage 2 respondents at each level of partisanship (1 through 7), and then averaging the favorability across respondents for each partisanship-advertisement combination.[^175a]

respondents. The distribution of prediction errors is consequential because it gives insight into the reliability of our estimator. The estimand stated in the previous section is the difference in outcome (Biden favorability, intent to vote Biden, intent to vote at all) between receiving the most effective advertisement and receiving any particular advertisement. If such a difference exists, the sensitivity of our estimator to this quantity depends on multiple factors: the size of the difference, the size of the sample, and *the difference between the true optimal schedule, $\mathbf{D}^*$ and the allocations made by the predictive model in the second stage, $\hat{\mathbf{D}}$*.

The optimal schedule is the schedule of advertisement that, given the schedule of stage 2 participants and their associated traits, assigns the individually optimal advertisement to each one:

$$\mathbf{D}^* = \{D^*(X_{i=1}), D^*(X_{i=2}), ...D^*(X_{i=900})\}$$

As calculating the optimal allocation $\mathbf{D}^*$ requires realizing mutually exclusive outcomes, it is not possible to know whether the allocation made in stage 2, $\hat{\mathbf{D}}$, was optimal, nor how much it differed from the true optimal allocation. However for each individual the fitted model generated a predicted outcome for each of the five advertisements, and one of these outcomes was realized. Using the realized outcomes, we can measure the actual prediction error of the targeting algorithm. To the extent that these predictions were accurate and unbiased, we can be more confident that the stage 2 allocations, $\hat{\mathbf{D}}$, are close to the true optimal schedule $\mathbf{D}^*$.

The overall error distribution and conditional error distributions exhibit a weak positive bias ($\mathbb{E}(e) = 0.21$, $p < 0.001$). This indicates that on average, the predictive algorithm slightly overestimated the effect of the advertisement. Looking at the individual conditional error distributions in panel 3, all distributions except Race have a mean significantly greater than zero ($p_{Race} = 0.36$). Finally, the long tail on the 2A ("Biden will come for your guns" in Table 1) and Race advertisements reflects that the predicted Biden favorability was nearest the extremes, resulting in a larger possible prediction error.

That the predictive algorithm exhibits a weak positive bias is not necessarily problematic. Within this design, the requirement of the algorithm was to be able to provide a correct ordinal ranking of the possible outcomes so that the most effective advertisement could be shown. The average prediction error associated with the first three advertisements—In-Group, Hunter Biden and 2A—are roughly similar (0.31, 0.28 and 0.34 respectively) and thus less likely to distort the ranking between each other.

Surprisingly, despite having the smallest bias, the Race advertisement was allocated to the greatest number of respondents (Panel 4). Panel 2 reveals why this is the case; this advertisement was more likely to be allocated to left-leaning respondents.[12] The bias associated with the Veterans advertisement suggests that the it may

[12] This matches intuitions about the advertisements: an advertisement attacking Biden's record on race issues is more likely

have been under-allocated. However, given that panel 4 shows that no single advertisement dominated the allocations, it does not appear that the bias was sufficiently large to create a scenario in which there was no targeting.

Sample Balance

In total, the experiment took place in a seven-hour window across five time zones (or three, excluding Alaska and Hawaii), with the switch-over between the first and second stage happening in under an hour. Causal interpretation of the results of this experiments relies on there not being systematic differences between the stage 1 (control) and stage 2 (treatment) samples. Because the two samples were not collected concurrently, it is not possible to rule out time-of-day effects. This section outlines some key randomization and balance checks to provide credence to the claims of successful causal identification made later in this article.

Chi-squared tests for statistical dependence between the treatment group indicator and pre-treatment covariates failed to reject the null hypothesis of independence for all pre-treatment covariates except for ideology, but this result was not robust to the Holm (1979) or Benjamini-Hochberg (1995) multiple comparisons corrections. This is consistent with the fact that dependence on Ideology but not Partisanship is theoretically unlikely.

A further check took the form of covariate balance tests and entropy balancing (Hainmueller 2012). The mean differences was calculated for all pre-treatment covariates other than age separated into dummy variables for each category, and for age the standardized mean difference was calculated. None of these differences exceeded 0.05.[13]

Micro-Targeting Effect

The effect of targeting was tested for three outcomes: (1) Biden favorability (2) intent to vote for Biden and (3) intent to vote at all. This section covers the key results, their implications for the hypotheses, and provides a brief outline of the robustness checks made. A full set of robustness checks along with the accompanying tables is available in Appendix II.

The key results of this study are in Figure 2. This figure shows the effect of targeting on Biden favorability, intent to vote for Biden, and intent to turnout among voters who had not already voted at the time of the survey, conditional on partisan self-identification. Each pair of bars in the figure shows the predicted level of each of the dependent variables for untargeted and targeted respondents, with 95% confidence intervals, for

to resonate with left-leaning voters than an advertisement about Second Amendment issues or Biden/Clinton mocking Trump supporters.

[13] See Figure 4, “Love” plot of mean differences, and Table 10: Model Fit with and without Entropy Balancing in *Appendix III: Implementation Notes.*

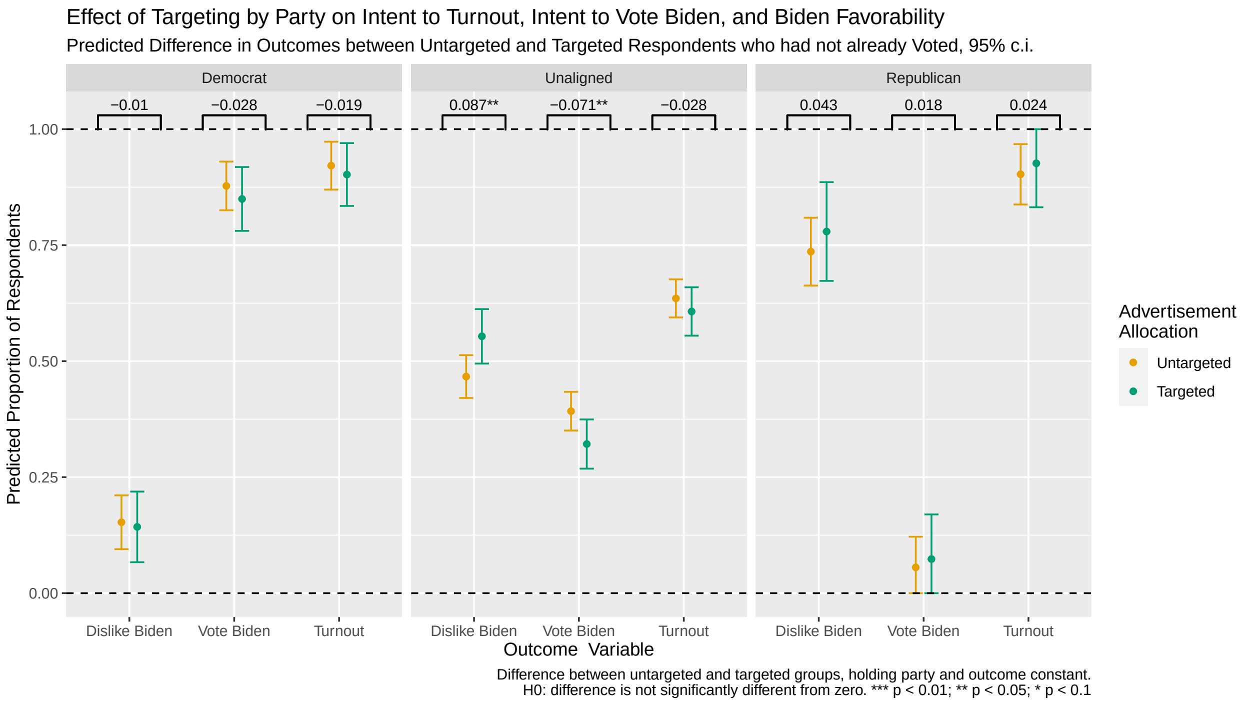

Figure 2: *levels and differences for effect of targeting on Biden favorability, intent to vote for Biden, and turnout, among respondents not pre-voting before survey, conditional on partisan (non-)affiliation.*

Democrat, Unaligned and Republican respondents in the corresponding facet.

The numbers along the top show the difference between the untargeted and targeted groups, which is equal to the coefficient on the treatment (targeting) in a regression with the corresponding group (Democrat, Unaligned, Republican) as the base partisanship category. The stars on the coefficients denote the p-value on a hypothesis test that this difference (i.e. the coefficient on treatment) is not significantly different from zero.

The differences and levels both reveal notable results. In the central panel, the difference between the untargeted and targeted groups for Biden favorability and intent to vote for Biden is 0.087 and -0.071 respectively. This translates to targeting causing a 8.7 percentage point *increase* in the proportion of unaligned respondents who state that they dislike Biden, and a 7.1 percentage point *decrease* in the proportion of unaligned respondents who state they intend to vote for Biden ($p < 0.05$).

The remainder of the differences are smaller in magnitude, ranging from -0.01 for the effect of targeting on Biden favorability among Democrats to 0.043 for the effect of targeting on Biden favorability for Republicans. Although not statistically significant, the differences provide evidence of a heterogeneous effect of targeting, with the sign on the coefficients flipped for Democrat and Republican voters.

The levels of the predicted proportions largely match intuitions. Respondents identifying as Democrat exhibit low levels of Biden dislike, high levels of intention to vote for Biden, and a high stated propensity to turnout to vote. Respondents identifying as Republican mirror this, with a high level of Biden dislike, low level of intent to vote for Biden, but a similarly high level of intent to turnout to vote. Unaligned voters show middling levels of Biden dislike, somewhat low levels of intent to vote Biden, and the lowest levels of intent to turnout to vote.

Models are also fitted estimating the effect of targeting on Biden favorability for the full sample of respondents, including those who had already cast their vote at the time of the survey. In the model fitting only treatment on Biden favorability without covariates or interactions, the effect of targeting is moderate and insignificant ($\beta = -0.06$, $s = 0.05$). Similarly, in the model interacting partisan self-identification on treatment, the coefficient on the treatment among unaligned respondents is large but insignificant ($\beta = -0.11$, $s = 0.07$).[14]

All results are checked for robustness with a variety of alternative specifications, not limited to the inclusion of pre-treatment covariates as controls, operationalization of the outcome as a five-point ordinal or continuous scale[15]. The key results on the effect of targeting on unaligned voters do not vary as a result of these checks, increasing confidence that the result is not dependent on a particular specification, operationalization, or otherwise spurious.

Discussion

Limitations

There are many senses in which the targeting as presented in this paper is not entirely realistic. For one, this experiment only shows that targeting can affect an individual's *stated* candidate favorability and intent to vote. Given this, and extensive evidence of temporal decay on advertising effects (Gerber et al. 2011; Mitchell 2012) the effect of targeting on actual vote choice is not explainable with the evidence gathered in this experiment.

The experiment also fails to replicate real-world conditions in numerous other ways. Typically voters will be continuously exposed to opposing messages from both campaigns, whereas in this experiment the effect is measured after viewing just one. On the other hand, because this experiment was run just prior to the election, it is likely that respondents were being exposed to a wide variety of advertisements outside of the context of the experiment[16].

As mentioned, this paper takes a different approach to micro-targeting from the psychometric profiling made

[14] Full results of these models are in in Table: 5, Appendix II.

[15] For the results of these tests and models, refer to Appendix II.

[16] Anecdotally, many of the respondents who reached out after the experiment stated that this was in fact the case.

famous by CA and tested by Zarouali et al. (2020). I argue that the psychometric approach relies on assertions about the congruity of psychological profiles with particular messages that are unnecessary because the outcome of interest can be directly optimized. Thus given a scenario where the psychometric approach and direct optimization approach are compared with the same information on voters, the optimization approach should in theory approach an upper bound of the effectiveness of a campaign with the same information. Put differently, even if the two approaches produce diverging results, the psychometric approach is unlikely to be more effective, and even if it is, the key point of this paper stands: *micro-targeting can work.*

That being said, the advantage of the psychometric approach and the reason that it will likely continue to be employed by campaigns is that the psychometric profiles can be used for both tailoring as well as targeting. The direct optimization strategy employed in this paper is only appropriate when the set of advertisements is already fixed.

Lastly, this experiment uses a convenience sample and is therefore not guaranteed to be completely representative. While the sample employed does sample roughly proportionately from all fifty states plus Washington D.C. (see Table 6, Appendix III) and has rough gender balance (Table 7, 50.2% female), it over-represents White and Asian respondents and under-represents Black and Hispanic respondents (Table 8). In order to simulate representative estimates, the sample was balanced to the nationally representative CCES 2019 sample (Ansolabehere, Schaffner, and Luks 2020) using entropy balancing (Hainmueller 2012), but due to the lack of elderly low-income respondents in my own sample, reasonable weights could not be produced (see Appendix III).

Significance

The results presented in the previous section provide strong evidence that the experimental design to emulate targeting was successful, and in turn that targeted advertising has a measurable effect among unaligned voters. Evidence in part 1 of the results supports the assertion that the RF model employed learned to predict respondents' potential outcomes under each of the five advertisements and used this information to match respondents with an individually more effective advertisement. Clearly, it is not possible to know whether the realized schedule of allocations is the true optimal one, $\mathbf{D} = \mathbf{D}^*$, but the information in Figure 1 presents evidence that the allocations were a) related to respondent characteristics in an expected way, and b) the model did not demonstrate systematic bias towards just one of the five advertisements.

Given this and the comparability of the two samples based on evidence presented, the difference between the treatment and control groups is due to the treatment group receiving individually optimized advertisements[17]—

[17] As noted in the section Sample Balance, given that the samples were collected at slightly different times of day—17:00 and

micro-targeting—and this effect is significant for unaligned voters. This raises two questions: what does it mean for targeting to have an effect, and what are the substantive implications of the estimated effect size?

The interpretation of the treatment in this experience differs from the more familiar case where treatment is exposure to some stimulus. Here, the treatment is the selection and allocation mechanism for the stimulus—the difference between seeing *some* advertisement, and *the most effective* advertisement. Thus the treatment effect does not indicate that receiving the stimulus by a different mechanism results in a difference in outcome, but that there is sufficient heterogeneity between the potential outcomes over the advertisements such that it is possible to detect a difference between the optimal outcome and the average one.

In other words, this experiment reveals that there is sufficient and substantial heterogeneity in the effect of advertisements on unaligned voters—a result that runs directly contrary to Coppock, Hill, and Vavreck (2020). I argue that these results are not as contrary as initially appears, however. Coppock, Hill, and Vavreck (2020) demonstrate in their experiments that there is little variation in the effectiveness of advertisements, but mostly search for heterogeneity on advertisement characteristics. Among respondent characteristics, the authors only test for heterogeneity on partisanship. In contrast, the design of the experiment employed in this paper leverages respondent heterogeneity across nine pre-treatment covariates, and then searches the space of 362880 possible combinations for the optimal outcome. The authors account for this possibility in their conclusion: *"[t]here may well be a mix of message features and subject characteristics that generates politically important persuasion"* (Coppock, Hill, and Vavreck 2020, 5)[18].

Having established the existence of an effect, the reader may be wondering what are the substantive implications of the estimated effect size. The experiment shows that in a hypothetical campaign where the Trump campaign ran just five advertisements, optimizing allocation with the method presented in this paper would have resulted in 8.7 percentage points more of unaligned voters stating that they do not like Biden, and 7.1 percentage points fewer of unaligned voters stating that they intend to vote for Biden, when compared to a scenario when advertisements are assigned independently of individual traits.

What do 8.7 and 7.1 percentage points mean? While it is not possible with the available data to give a rigorous estimate, the following calculation gives an illustration of how meaningful a 7.1 percentage point swing could be. In the 2020 United States presidential election, 31 states included information on partisan affiliation for voter registration. Assuming (unrealistically) that all registered voters turn out to vote, that stated intention not to vote for Biden translates into voting for neither candidate (i.e. no switching), then we can multiply the proportion of unaligned voters in each state by the effect size in order to estimate how the

19:30 GMT on 28 October 2020—the difference may also be due to a temporal effect.

[18] The links between this paper and Coppock, Hill, and Vavreck (2020) are discussed in greater detail in Appendix I.

result would have changed.

For example, in Florida, where 26.7% of registered voters affiliated with neither of the major parties, a swing of 7.1 percentage points among unaligned voters could lead to a diminution of $0.267 \times 0.0708 = 0.0189$, or 1.9 percentage points in the Biden vote share. In North Carolina and Arizona, this estimated change is larger than the percentage point difference between the shares of the two leading candidates (in Arizona, 35.1% unaligned with a 0.3% margin of victory, and in North Carolina 30.6% of voters are unaligned with a 1.3% margin of victory).

There are many obvious caveats and limitations to this calculation beyond the unrealistic assumptions already stated. For one, changing answers on a survey is radically different to changing voting intention. For another, this survey design and calculation do not account for decay and counter-acting effects, meaning the actual effect is likely smaller. A further point is that the 7.1 percentage point effect is based on a distribution of Biden preferences that were present in the sample; this is likely not the same for all possible distributions of starting preferences. Finally, the sample used in this survey is not representative[19], meaning that the distribution of effects will likely be different. Nevertheless, this calculation is meant to illustrate that the magnitude of effect is not trivial in a political context like the United States, where consequential races have incredibly small margins.

The non-triviality of this technology highlights the potential for a multitude of harms. For one, this technology creates incentives to disregard voter privacy, which is arguably a harm in of itself (Wachter 2017). Moreover, because those using this technology are more likely to win power, it creates a self-reinforcing cycle (Hindman 2018, 5) in which private actors (Zuboff 2015; Baldwin-Philippi 2017) and public officials (Krotoszynski 2020) broker access to such data.

There is a second, less clear, set of normative issues that speaks to manipulation and the legitimacy of democratic processes. It begins with asking "what does it mean to show a voter the *right* advertisement?" Proponents of targeted advertising often present it in a benign and efficiency-improving frame; targeted advertising reduces the informational burden by showing the consumer only the goods that are most relevant to their interests (Hindman 2018, 39). Analogously, targeted political advertising has the potential for good; it can be used to show voters elements of candidates' policies that are most relevant to their interests, lowering the cost of civic and political engagement.

Targeted advertising can also be malevolent. Combining emotive advertisements with knowledge of the fears and anxieties of the voters[20], targeted negative advertising could fairly be called a form of manipulation

[19] See *Appendix III: Implementation Notes* for notes on the representativeness of the sample employed in this study.

[20] Note that with the black box algorithmic approach to targeting employed here, the advertisers may not specifically know

(Sunstein 2015). By seeking to identify ways to elicit strong negative reactions from voters, campaigns are attempting to engage peripheral, or emotive, and not central, or cognitive/logical, processing of the information in advertisement. This approach to advertising attempts to bypass the deliberative and cognitive capacity of the viewer, ultimately seeking to make the decision for the viewer.

This negative and manipulative strategy is a more suitable description of the calculus employed by CA and demonstrated in the five Trump campaign advertisements used in this experiment. The psychometric profiling employed by CA targeted “neurotic” voters that would be especially susceptible to fear-based advertising (Hindman 2018). The five advertisements similarly appear to be attempting to elicit negative emotions to be associated with Biden, from the fear that he will take away the viewer’s guns, to emphasizing that Biden and Clinton laugh at and mock people like the viewer. To determine the extent to which this strategy characterizes the various United States campaigns requires further research.

However, these are issues with manipulation and not targeting. At an individual level, manipulative political advertising could be seen as disenfranchising, by depriving voters of the opportunity to make their own decision. The results from this experiment show that such manipulation is possible on a massive scale, and can therefore affect the outcome of an election. That election outcomes could be determined by micro-targeted manipulative advertisement campaigns undermines the legitimacy derived from the democratic process, as such outcomes arguably no longer reflect the public will. In a nutshell, although campaigns can engage in manipulative practices without targeting, to the extent that a manipulative campaign is effective at a societal level the threat transcends from individual to systemic.

There are a number of clear follow-ups to clarify the scope of these conclusions and further explore the normative claims made. The first follow-up takes the form of an experiment with larger N and more treatments, in the form of a brief being shown prior to the advertisement to inform the respondent that the advertisement was targeted to them. It may be the case that knowledge of being targeted nullifies any effect the advertisement may have, or even reverse it[21]. In addition to testing these hypotheses, this follow-up will help assess the extent to which the patterns of effects found in this paper are specific to the five advertisements employed.

Finally, in order to be able to make stronger normative claims, we need to know more about the modes of persuasion being used by the various campaigns. As noted, targeting may not be problematic if seen as a way to increase efficiency and engagement. The manipulative aspect appears when elements of a respondent’s background and psychology are turned against them, to prevent them from cognitively engaging with the

what aspects of an advertisement make it resonate with an individual. This does not mean, however, that the algorithm is not implicitly leveraging these patterns or identifying these psychological traits.

[21] Kruikemeier, Sezgin, and Boerman (2016) test the effect of knowledge of being targeted, but do not in fact target the advertisements, leaving open the possibility that respondents may be told they are being targeted and then shown an incongruous advertisement.

information being presented. Thus in order to claim whether targeting was used problematically, the next step is to characterize the variety of strategies employed by the various campaigns.

References

Ansolabehere, Stephen, and Shanto Iyengar. 1995. *Going Negative : How Attack Ads Shrink and Polarize the Electorate*. New York ; London: Free Press.

Ansolabehere, Stephen, Brian Schaffner, and Samantha Luks. 2020. "CCES Common Content, 2019." Harvard Dataverse. https://doi.org/10.7910/DVN/WOT7O8.

Baldwin-Philippi, Jessica. 2017. "The Myths of Data-Driven Campaigning." *Political Communication* 34 (4): 627–33.

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society: Series B (Methodological)* 57 (1): 289–300.

Broockman, David E, and Donald P Green. 2014. "Do Online Advertisements Increase Political Candidates' Name Recognition or Favorability? Evidence from Randomized Field Experiments." *Political Behavior* 36 (2): 263–89.

Burkell, Jacquelyn, and Priscilla M. Regan. 2019. "Voter Preferences, Voter Manipulation, Voter Analytics: Policy Options for Less Surveillance and More Autonomy." *Internet Policy Review* 8 (4).

———. 2020. "Voting Public: Leveraging Personal Information to Construct Voter Preference." In *Big Data, Political Campaigning and the Law: Democracy and Privacy in the Age of Micro-Targeting*, edited by Normann Witzleb, Moira Paterson, and Janice Richardson, 47–68. Routledge.

Cialdini, Robert B. 2007. *Influence: The Psychology of Persuasion*. Vol. 55. Collins New York.

Coppock, Alexander, Seth J Hill, and Lynn Vavreck. 2020. "The Small Effects of Political Advertising Are Small Regardless of Context, Message, Sender, or Receiver: Evidence from 59 Real-Time Randomized Experiments." *Science Advances* 6 (36).

Dobber, Tom, Ronan Ó Fathaigh, and Frederik Zuiderveen Borgesius. 2019. "The Regulation of Online Political Micro-Targeting in Europe." *Internet Policy Review* 8 (4).

Edelson, Laura, Shikhar Sakhuja, Ratan Dey, and Damon McCoy. 2019. "An Analysis of United States Online Political Advertising Transparency." *arXiv Preprint arXiv:1902.04385*.

Fowler, Erika Franklin, Michael M Franz, Gregory J Martin, Zachary Peskowitz, and Travis N Ridout. 2021. "Political Advertising Online and Offline." *American Political Science Review* 115 (1): 130–49.

Gerber, Alan S, James G Gimpel, Donald P Green, and Daron R Shaw. 2011. "How Large and Long-Lasting

Are the Persuasive Effects of Televised Campaign Ads? Results from a Randomized Field Experiment." *American Political Science Review* 105 (1): 135–50.

Grimmer, Justin, Margaret E. Roberts, and Brandon M. Stewart. 2021. "Machine Learning for Social Science: An Agnostic Approach." *Annual Review of Political Science* 24 (1): null. https://doi.org/10.1146/annurev-polisci-053119-015921.

Hainmueller, Jens. 2012. "Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies." *Political Analysis*, 25–46.

Hindman, Matthew. 2018. *The Internet Trap.* Princeton University Press. https://doi.org/10.1515/9780691184074.

Holm, Sture. 1979. "A Simple Sequentially Rejective Multiple Test Procedure." *Scandinavian Journal of Statistics*, 65–70.

Kang, Michael S. 2005. "From Broadcasting to Narrowcasting: The Emerging Challenge for Campaign Finance Law." *George Washington Law Review* 73 (5-6): 1070–95.

Krotoszynski, Ronald J., Jr. 2020. "Big Data and the Electoral Process in the United States: Constitutional Constraint and Limited Data Privacy Regulations." In *Big Data, Political Campaigning and the Law: Democracy and Privacy in the Age of Micro-Targeting*, 186–213. Routledge.

Kruikemeier, Sanne, Minem Sezgin, and Sophie C Boerman. 2016. "Political Microtargeting: Relationship Between Personalized Advertising on Facebook and Voters' Responses." *Cyberpsychology, Behavior, and Social Networking* 19 (6): 367–72.

Madsen, Jens Koed, and Toby D Pilditch. 2018. "A Method for Evaluating Cognitively Informed Micro-Targeted Campaign Strategies: An Agent-Based Model Proof of Principle." *PloS One* 13 (4).

Mitchell, Dona-Gene. 2012. "It's about Time: The Lifespan of Information Effects in a Multiweek Campaign." *American Journal of Political Science* 56 (2): 298–311.

Montgomery, Jacob M., and Santiago Olivella. 2018. "Tree-Based Models for Political Science Data." *American Journal of Political Science* 62 (3): 729–44. https://doi.org/10.1111/ajps.12361.

Nickerson, David W., and Todd Rogers. 2020. "Campaigns Influence Election Outcomes Less Than You Think." *Science* 369 (6508): 1181–82. https://doi.org/10.1126/science.abb2437.

Pariser, E. 2011. *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think*. Penguin Publishing Group.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, et al. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12: 2825–30.

Petty, Richard E, and John T Cacioppo. 1986. "The Elaboration Likelihood Model of Persuasion." In *Communication and Persuasion*, 1–24. Springer.

Simon, Felix M. 2019. "'We Power Democracy': Exploring the Promises of the Political Data Analytics Industry." *The Information Society* 35 (3): 158–69.

Sunstein, Cass R. 2015. "Fifty Shades of Manipulation."

Wachter, Sandra. 2017. "Privacy: Primus Inter Pares — Privacy as a Precondition for Self-Development, Personal Fulfilment and the Free Enjoyment of Fundamental Human Rights." *SSRN*, January.

Wood, Abby K, and Ann M Ravel. 2017. "Fool Me Once: Regulating Fake News and Other Online Advertising." *S. Cal. L. Rev.* 91: 1223.

Zarouali, Brahim, Tom Dobber, Guy De Pauw, and Claes de Vreese. 2020. "Using a Personality-Profiling Algorithm to Investigate Political Microtargeting: Assessing the Persuasion Effects of Personality-Tailored Ads on Social Media." *Communication Research.*

Zuboff, Shoshana. 2015. "Big Other: Surveillance Capitalism and the Prospects of an Information Civilization." *Journal of Information Technology* 30 (1): 75–89.

Appendix I: Theoretical Notes

Coppock, Hill, and Vavreck (2020)

This paper makes extensive reference to Coppock, Hill, and Vavreck (2020) (within this section of the appendix, CHV20). CHV20 and its replication materials, published just over a month prior to this experiment, were essential to its design and testing. The replication data was used as mock results for the first stage, which were then used to train predictive algorithms and simulate the targeting in stage two.

The simulated targeting experiment based on the CHV20 data indicated that targeting would have an effect, with a magnitude of roughly 0.4 on a 1-5 scale. This result, however, was difficult to interpret because the expected outcome under targeting was given by the same model that was used to optimize advertisement allocations. To increase certainty that this result was not an artifact of the algorithm selectively sampling off the right-hand-side of the standard error, I conducted a permutation test ($n = 30,000$) in which the treatment vector was randomized. The null hypothesis being tested by this permutation test was row-level treatment independence, which would make the targeting irrelevant. This null hypothesis was rejected with $p = 0.99$.

I subsequently used the estimated effect size and variance as the basis for a pre-experimental power analysis, shown in Figure 3. On this basis I removed two additional treatment categories, which I intend to test in a follow-up experiment.

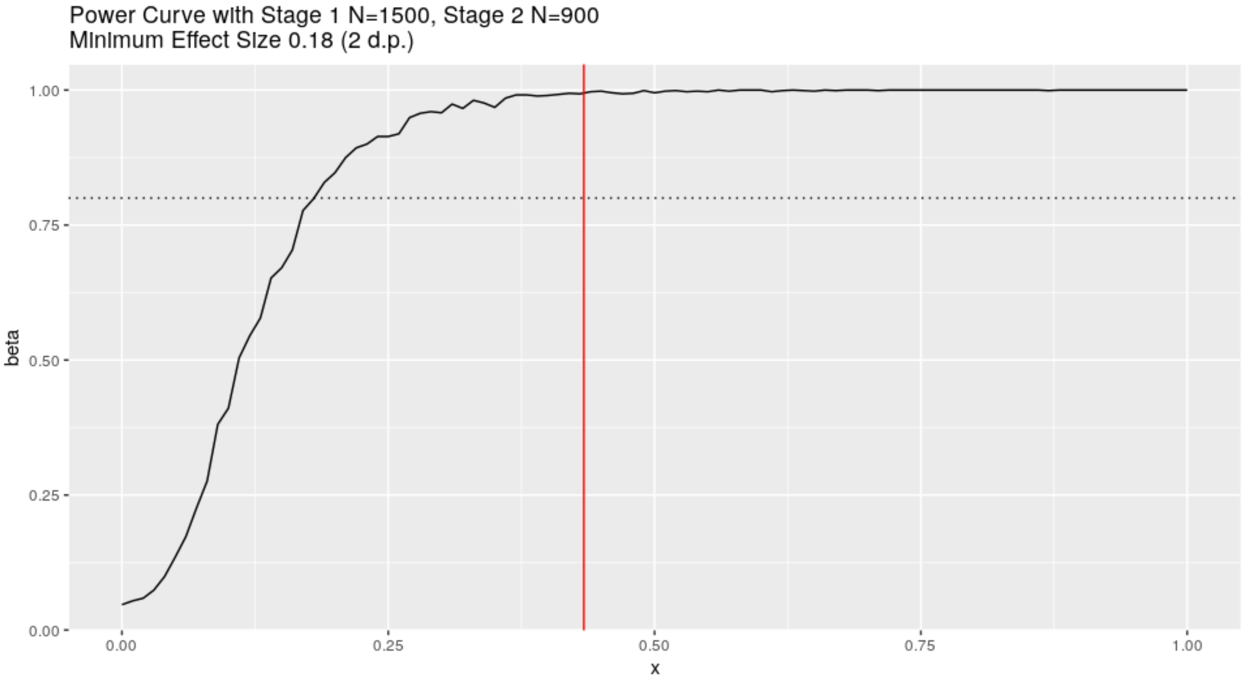

Figure 3: *power analysis from CHV20 data*

As noted, the conclusions of this paper seem at odds with CHV20, who find little evidence of treatment effect heterogeneity. They infer from this that there is little room for micro-targeting to work; if the effect of advertisements does not vary greatly from one individual to another, why should micro-targeting make any difference?

I suggest that there are two main reasons for this difference. The first is that whereas CHV20 focuses primarily on heterogeneity over advertisements, this experiment maximizes heterogeneity within voters. Thus, although CHV20 search for conditional effects over the characteristics effect of 49 different advertisements, the only respondent characteristic they test for heterogeneity over is partisanship.

In comparison, the targeting algorithm trained for this experiment searches a space of up to $362,880$ combinations of respondent characteristics to find connections between response to the advertisement and combinations of traits. The respondent is then shown the advertisement predicted to maximize (or minimize) this difference. As referenced in the main body, CHV20 does indeed note that exhaustively searching traits may expose sources of heterogeneity.

Average Effects for Stage 1 Advertisements

A theoretical quantity of interest is the effect of each individual advertisement used in this experiment. Randomized advertisement assignment in the first stage allows for a causally identified interpretation of the coefficient on each advertisement, but because it is unclear which advertisement should serve as a reference category, these coefficients are not meaningful. The original design for the experiment included a control advertisement in stage 1 in order facilitate these comparisons, but this was omitted to prioritize the key components of this experiment: maximizing the number of observations for training the predictive algorithm and identifying the effect of targeting. The implementation of the control group can be seen in the source code of this project on GitHub.

Appendix II: Regression Results and Robustness Checks

The following appendix contains the full results of the various models and operationalizations, as well as various robustness checks.

Effect on Candidate Favorability

2 reports the effect of targeting conditional on partisanship for the subset of respondents who had not already voted at the time of the survey ($N = 1,160$). The coefficients in this table reveal a substantial amount of treatment effect heterogeneity. The second row of coefficients, *Targeting (Unaligned)*, shows the effect of targeting on respondents who had not already cast their votes and self-identify as neither party; the group that a campaign would aim to target. The estimated CATE is -0.218 points on a five-point scale ($SE = 0.093$), *which translates to an* 8.7 *percentage point increase in respondents saying they dislike Biden* ($SE = 3.8$ p.p). Both of these coefficients are significantly different from zero at standard confidence levels of $\alpha = 0.05$ ($p = 0.0192,\ 0.0338,\ 0.0228,\ 0.0416$).

Table 2: Effect of Micro-Targeting on Candidate Favorability, Interacted on Partisan Self-Identification among Respondents who had not Voted

	OLS: Five-Point	Ordered Logistic	OLS: Binary	Logistic
Intercept	2.588^{***}		0.467^{***}	-0.133
	(0.057)		(0.024)	(0.105)
Targeted (Unaligned)	-0.218^{**}	-0.318^{**}	0.087^{**}	0.348^{**}
	(0.093)	(0.150)	(0.038)	(0.171)
Democrat	0.997^{***}	1.625^{***}	-0.314^{***}	-1.580^{***}
	(0.092)	(0.159)	(0.038)	(0.212)
Republican	-0.693^{***}	-1.276^{***}	0.269^{***}	1.159^{***}
	(0.108)	(0.189)	(0.044)	(0.216)
Targeted × Democrat	0.362^{**}	0.640^{**}	-0.097	-0.427
	(0.151)	(0.251)	(0.062)	(0.353)
Targeted × Republican	0.263	0.352	-0.043	-0.112
	(0.186)	(0.320)	(0.076)	(0.388)
R^2	0.258		0.190	
Adj. R^2	0.254		0.187	
Num. obs.	1160	1160	1160	1160
AIC		3259.343		1363.093
BIC		3304.849		1393.430
Log Likelihood		-1620.671		-675.547
Deviance		3241.343		1351.093

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

The remaining coefficients reveal unsurprising patterns. The effect of self-identifying as Democrat and Republican has large and significant effects on Biden favorability in the expected directions. Looking at the CATEs of self-identification as Democrat or Republican, we discover that their coefficients are in the same direction and diminish the effect of targeting. In other words, targeting anti-Biden advertisements has the strongest effect on unaligned voters, but has a relatively weak effect on voters who identify with a party.

Effect on Voting Preference

The dependent variable for the models in this section is the proportion of respondents stating their intention to vote for Biden in the general election, out of the respondents who had not voted at the time of the survey. 3 reports the effect of targeting on intention to vote for Biden among respondents who had not voted at the time of the survey. The two columns on the left report the models regressing targeting on voting preference, and the two columns on the right report the models regressing targeting interacted on partisan self-identification on voting preference. The columns alternatingly report the results for OLS and logistic regression models.

Table 3: Effect of Micro-Targeting on Intention to Vote for Biden among Respondents who had not Voted

	OLS Uninteracted	Logit Uninteracted	OLS Interacted	Logit Interacted
Intercept	0.478^{***}	-0.090	0.392^{***}	-0.438^{***}
	(0.018)	(0.074)	(0.021)	(0.108)
Targeted (Unaligned)	-0.030	-0.123	-0.071^{**}	-0.309^{*}
	(0.030)	(0.122)	(0.034)	(0.179)
Democrat			0.485^{***}	2.409^{***}
			(0.034)	(0.229)
Republican			-0.337^{***}	-2.395^{***}
			(0.040)	(0.379)
Targeted × Democrat			0.043	0.070
			(0.056)	(0.363)
Targeted × Republican			0.089	0.609
			(0.069)	(0.617)
R^2	0.001		0.346	
Adj. R^2	0.000		0.343	
Num. obs.	1160	1160	1160	1160
AIC		1605.846		1158.461
BIC		1615.958		1188.798
Log Likelihood		-800.923		-573.230
Deviance		1601.846		1146.461

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

The first pair of models (1) and (2) show a weak, insignificant and negative ATE of targeting on voting preference. When we search for heterogeneity over (non-) partisanship, we observe a similar pattern to the CATE of targeting on candidate favorability. The effect of targeting on intention to vote for Biden is -0.071 ($SE = 0.034$, $p = 0.03967$), meaning that among respondents who identified with neither party and had not voted at the time of the survey, being targeted increased the proportion of respondents not intending to vote for Biden by 7.1 percentage points. As with the previous section, the effect of aligning with either the Democratic or Republican party largely nullifies the effect of targeting. That the pattern is persisting in a separate but related outcome increases confidence that targeting is in fact increasing the likelihood that the targeted advertisements are persuasive.

Effect on Turnout

Table 4: Effect of Micro-Targeting on Turnout among Respondents who had not Voted

	OLS Uninteracted	Logit Uninteracted	OLS Interacted	Logit Interacted
Intercept	0.777***	1.250***	0.628***	0.523***
	(0.015)	(0.086)	(0.020)	(0.105)
Targeted (Unaligned)	−0.021	−0.118	−0.019	−0.082
	(0.025)	(0.139)	(0.033)	(0.170)
Democrat			0.298***	2.002***
			(0.032)	(0.267)
Republican			0.285***	1.821***
			(0.037)	(0.299)
Targeted × Democrat			0.004	−0.118
			(0.053)	(0.416)
Targeted × Republican			0.035	0.303
			(0.065)	(0.568)
R^2	0.001		0.126	
Adj. R^2	−0.000		0.123	
Num. obs.	1241	1241	1241	1241
AIC		1343.142		1184.504
BIC		1353.389		1215.246
Log Likelihood		−669.571		−586.252
Deviance		1339.142		1172.504

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

The final set of results, shown in 4, looks at the effect of targeting on turnout on respondents who had not voted yet. This is operationalized as a dummy variable indicating the proportion of respondents stating that they will vote for Biden, Trump, or a third candidate. The models are presented in the same as the previous section, with the first two columns testing an uninteracted model and the latter two testing a model interacting turnout intention on partisan self-identification.

These models indicate that if there is an effect of targeting on turnout, then it is not significantly different from zero in the total sample nor among non-partisan voters. It is worth noting that partisan voters were more likely to indicate that they intended to vote by 28.6 and 26.7 percentage points for Democrats and Republicans respectively, from a baseline of 63.5% for non-partisan voters.

Biden Favorability with Full Sample

Unlike the other two outcomes, the effect of targeting on Biden favorability could be tested on the full sample of respondents. The results are reported in Table 5.

Table 5: Effect of Micro-Targeting on Candidate Favorability Interacted on Partisan-Self Identification

	Uninteracted Model		Interacted on Partisan Self-Identification	
	(1) Linear	(2) Binary	(3) Linear	(4) Binary
Targeted (Unaligned)	-0.06	0.03	-0.11	0.05^{**}
	(0.05)	(0.02)	(0.07)	(0.03)
Democrat			1.10^{***}	-0.32^{***}
			(0.06)	(0.02)
Republican			-0.89^{***}	0.33^{***}
			(0.09)	(0.03)
Targeted × Democrat			0.08	-0.05
			(0.10)	(0.04)
Targeted × Republican			0.23	-0.07
			(0.14)	(0.06)
R^2	0.00	0.00	0.31	0.23
Adj. R^2	0.00	0.00	0.31	0.23
Num. obs.	2261	2261	2261	2261

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Appendix III: Implementation Notes

Operationalization Decisions

In the previous section of the appendix, the various regression models reflect different methods of operationalizing the outcome. As these decisions should not, and were not made arbitrarily, I explain the details and rationale here. For the candidate favorability question, after watching the advertisement, survey respondents were posed the following question:

> *"On a scale of 1-5, how would you rate Democratic Presidential Candidate Joseph Biden? (A rating of 4 or 5 means that you feel favorable and warm towards the candidate. A rating of 1 or 2 means that you don't feel favorable and warm towards the candidate. You would rate them at 3 if you don't feel particularly warm or cold toward them.)"*

This item is operationalized in four ways. The first two keep the five categories, but treat it either as a continuous linear scale, or as an ordinal categorical scale. The latter two treat it as a binary indicator of Biden dislike, with 1-2 being coded as 0, and 3-5 being coded as 1. The direction of the measure and inclusion of 3 in the positive category are not arbitrary; Biden *dislike* should be of particular interest to campaigns. If they can get voters to actively dislike the candidate (1-2), then this is more significant than simple indifference (3).

The vote choice question simply allowed them to indicate whether they intended to, or had already voted for Trump, Biden, another candidate, that they did not intend to vote, or that they did not wish to answer. The intent to vote Biden variable is simply a binary indicator on whether they chose "I intend to vote for Biden" for this question.

Representativeness

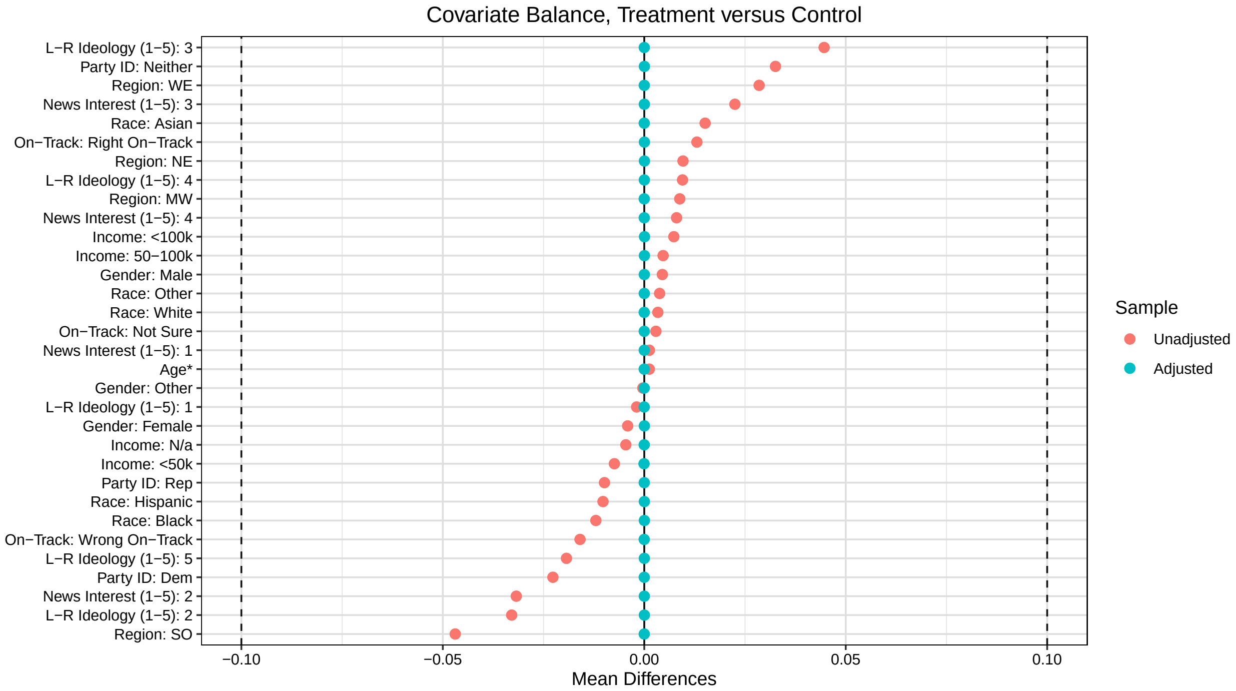

Figure 4: Covariate Balance on Treatment Indicator

As mentioned, due to resource constraints it was not possible to procure a representative sample. Nevertheless, the sample provided by Prolific covered all 50 states and Washington DC, with a roughly proportional number of respondents to the population of each state (6).

Table 6: Number of respondents per state (or district), with percentage over/under representation.

State	N	Distortion	State	N	Distortion
Alabama	17	0.74%	Montana	5	0.1%
Alaska	3	0.09%	Nebraska	12	0.06%
Arizona	53	-0.13%	Nevada	31	-0.43%
Arkansas	10	0.48%	New Hampshire	6	0.15%
California	307	-1.54%	New Jersey	79	-0.79%
Colorado	33	0.29%	New Mexico	8	0.28%
Connecticut	15	0.42%	New York	189	-2.43%
Delaware	10	-0.15%	North Carolina	78	-0.25%

State	N	Distortion	State	N	Distortion
District of Columbia	8	-0.14%	North Dakota	1	0.19%
Florida	185	-1.64%	Ohio	73	0.33%
Georgia	69	0.18%	Oklahoma	15	0.54%
Hawaii	10	-0.01%	Oregon	38	-0.4%
Idaho	7	0.23%	Pennsylvania	91	-0.12%
Illinois	105	-0.78%	Rhode Island	6	0.06%
Indiana	42	0.19%	South Carolina	28	0.33%
Iowa	15	0.3%	South Dakota	2	0.18%
Kansas	14	0.27%	Tennessee	36	0.49%
Kentucky	27	0.17%	Texas	175	1.09%
Louisiana	18	0.62%	Utah	7	0.67%
Maine	12	-0.12%	Vermont	4	0.01%
Maryland	54	-0.55%	Virginia	73	-0.63%
Massachusetts	58	-0.47%	Washington	54	-0.07%
Michigan	54	0.65%	West Virginia	6	0.28%
Minnesota	25	0.61%	Wisconsin	43	-0.13%
Mississippi	12	0.38%	Wyoming	3	0.04%
Missouri	35	0.32%			

Randomization Check and Time-of-Day Effects

Using the CCES data (Ansolabehere, Schaffner, and Luks 2020) and entropy balancing (Hainmueller 2012), I also attempted to simulate the results after adjusting covariate balance to a nationally representative sample.

The causal identification of the treatment relies on there not being any systematic differences between the treatment (targeted) and control (untargeted) groups. That the control data had to be gathered prior to the treatment step leaves open the possibility of bias due to the difference in time of day. In order to mitigate this bias, the entire experiment was run in as small a window as possible. In total, the experiment took place in a seven-hour window, with the switch-over between the first and second stage occurring in under an hour. The results of Chi-Squared tests of independence on assignment to treatment or control against all of the pre-treatment covariates are presented in Table: 11. All hypotheses fail to achieve significance except for ideology, but this is not robust to the Holm (1979) or Benjamini-Hochberg (1995) multiple comparisons

Table 7: Age and Gender Representation

Age		Census			Sample		
>=	<	Both	Male	Female	Both	Male	Female
15	19	8.09%	8.39%	7.81%	3.23%	3.41%	2.77%
20	24	8.25%	8.53%	7.98%	21.3%	21.3%	20.5%
25	29	9.03%	9.38%	8.7%	20.1%	19.4%	20.6%
30	34	8.51%	8.7%	8.33%	19.6%	21.6%	17.8%
35	39	8.32%	8.46%	8.19%	13.1%	14.7%	12%
40	44	7.6%	7.66%	7.54%	8.25%	9.05%	7.76%
45	49	7.9%	7.95%	7.84%	5.41%	4.82%	6.1%
50	54	7.9%	7.9%	7.9%	3.79%	2.23%	5.43%
55	59	8.21%	7.99%	8.42%	2.06%	1.29%	2.88%
60	64	7.99%	7.81%	8.16%	1.62%	0.823%	2.44%
65	69	6.74%	6.52%	6.94%	1.06%	1.18%	0.998%
70	74	5.48%	5.32%	5.64%	0.335%	0.118%	0.554%
75	79	3.63%	3.37%	3.88%	0.112%	0.118%	0.111%
80	84	2.35%	2%	2.68%	0%	0%	0%

Table 8: Race and Ethnic Representation

Race	Census (2019)	Sample
White	60.1%	66.29%
Black	13.4%	9.16%
Hispanic	18.5%	7.03%
Asian	5.9%	12.8%

Table 9: Income Bound Representation

Income Bound	Census (2019)	Sample
$0 to $53.5k	40%	54.67%
$53.6k to $109.7k	30%	29.4%

Table 10: Model Fit with and without Entropy Balancing

	Full Sample		Non Pre-Voting Subset	
	Unbalanced	Balanced	Unbalanced	Balanced
Targeted (Unaligned)	0.05^{**}	0.04	0.09^{**}	0.08^{*}
	(0.03)	(0.03)	(0.04)	(0.04)
Democrat	-0.32^{***}	-0.33^{***}	-0.31^{***}	-0.31^{***}
	(0.02)	(0.02)	(0.04)	(0.04)
Republican	0.33^{***}	0.31^{***}	0.27^{***}	0.27^{***}
	(0.03)	(0.04)	(0.04)	(0.05)
Targeted × Democrat	-0.05	-0.04	-0.10	-0.10
	(0.04)	(0.04)	(0.06)	(0.06)
Targeted × Republican	-0.07	-0.05	-0.04	-0.04
	(0.06)	(0.06)	(0.08)	(0.08)
R^2	0.23		0.19	
Adj. R^2	0.23		0.19	
Num. obs.	2261	2261	1160	1160
Deviance		390.06		233.43
Dispersion		0.17		0.20

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

corrections. I therefore conclude that there was a successful randomization, but for each model I additionally test a variant controlling for all pre-treatment covariates. These are reported in the main body of the article.

Table 11: Chi-Squared Test of Treatment on Pre-Treatment Independence, with Holm and Benjamini-Hochberg corrections.

	p	Holm	BH
Age	0.345066	1	0.7111003
Gender	0.9753692	1	0.9753692
Race	0.5646555	1	0.8873158
Income	0.9483788	1	0.9753692
Region	0.234541	1	0.7111003
NewsInt	0.2219223	1	0.7111003
On-Track?	0.9516303	1	0.9753692
Party	0.3878729	1	0.7111003
Ideology	0.02123161	0.2335477	0.2335477
General Vote	0.7472937	1	0.9753692

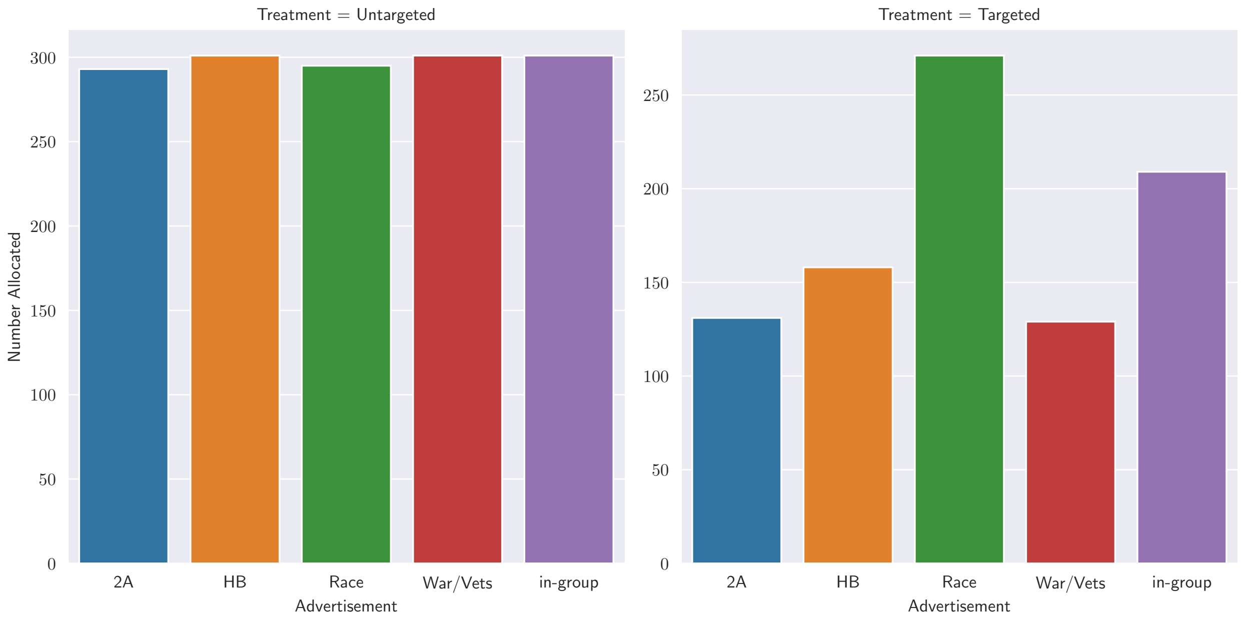

Figure 5: Advertisement Allocations

A second randomization check takes the form of the advertisement assignments. In the first stage, advertisements were assigned randomly using permuted block randomization. In the second stage, the advertisement predicted by the on-line random forest model to have the strongest effect was assigned. Below, Figure 5 shows the results of this randomization. In stage one, all advertisements are roughly equal, whereas in stage two the "Race" advertisement is the most probable. That the stage two allocation is not uniform or entirely placed in one advertisement is reassuring: the former might indicate that the algorithm was as good as random, whereas the latter would indicate that one advertisement out-performed all others, in which case micro-targeting is pointless. Note that to preclude this possibility, the five advertisements used were explicitly chosen with the aim of each having a specific and narrow target audience.

Survey Implementation

The design of this survey required a high degree of interactivity. In the second stage, the respondents' answers were sent to an on-line pre-trained machine learning model that would respond with the optimal advertisement assignment in real time. Given that there was no commercial survey software that provided this integration functionality, the survey website was built by the researcher from the ground up.

The front-end of the website[22] was largely written in PHP and hosted on an AWS Lightsail instance running a Linux-Nginx-MariaDB-PHP (LEMP) stack. PHP was likewise used to communicate with the back-end in

[22] CSS and other styling elements were copied then modified from `https://surveyjs.io/`.

real-time, which consisted of a Jupyter kernel and MariaDB database.

The machine learning algorithm was implemented in Python using the scikit-learn library (Pedregosa et al. 2011). Due to severe resource constraints on the Lightsail server, the algorithm was trained during the switch-over on a local machine, and the trained model was stored as a `joblib` binary then uploaded via `ssh` connection.

Interactivity between PHP and the algorithm was implemented using a modified Jupyter interface, which also handled queue management. Prediction response time during peak loads never exceeded 100ms.

The survey can be viewed at `https://survey.polinfo.org`.

The source code for the website is hosted on Github at `https://github.com/muhark/dotas-design`.

Advertisements

The five advertisements were selected from nearly one hundred videos with a length between 15 and 35 seconds posted by the Trump campaign to their YouTube channel in the final five months of the 2020 United States presidential contest. These were downloaded using the `youtube-dl` tool and hosted on the survey website listed above.

After narrowing down the videos to 10 likely candidates, I asked a panel of ten doctoral students at the University of Oxford to help select five advertisements based on the criteria that:

- The advertisements were clearly tailored to different audiences.
- No one advertisement was likely to outperform all others for all respondents.

Irregularities and Attention Check

Responses that failed one of two checks have been omitted from the data used for this article. Prolific provides basic demographic data on respondents that can be downloaded after respondents have completed the survey. Responses where there were considerable discrepancies between answers and supplied demographic information were rejected.

There was also a attention check immediately after the advertisement, which asked respondents which campaign ran the advertisement ("My name is ______ and I approve of this message"). Given that the answer was provided in the last few seconds of the advertisement, and the question was asked less than a few seconds later, I assumed that respondents who failed this were not paying attention to the video and therefore rejected their responses from the final data.

Reproduction Material

Reproduction code will be hosted on github at `https://github.com/muhark/dotas-design`. Data will be made available once appropriately sanitized and the period stated in the consent form for revoking consent has expired.

Ethics

Ethical approval for this experiment was applied for and obtained from the University of Oxford Central University Research Ethics Committee (CUREC) Department of Politics and International Relations Departmental Research Ethics Committee (DREC), with Research Ethics Approval Reference Number `SSH_DPIR_C1A_20_019`.

Prior to the experiment, participants were told what data would be collected, how it would be used, and provided various methods that they might withdraw their consent without consequence at any point in time. At the end of the experiment, participants were given the following debrief:

- **What was the purpose of this experiment?**

  The purpose of this experiment was to show whether "micro"-targeted political campaigning makes a difference to election outcomes. If it does, I plan to use this evidence to make the case for clearer disclosure requirements on advertisers.
- **How does this study show that?**

  During this experiment, you were assigned to either the "control" or "treatment" group. The control group was shown a random advertisement. The treatment group was given the "optimal" ad based on their answers to the first set of questions, chosen by an machine learning algorithm. By looking at the difference in average outcome between these two groups, we can make claims about "effect" of being targeted.
- **I want to know more!**

  If you have further questions about the survey, payment, want to hear about the research when it is published, or change your mind regarding consent, please contact me in the email address provided in the Participant Information Sheet.

Participants were compensated at an average rate of £8.86 per hour, depending on the length of time participants required.

Funding

This study was made possible from the researcher's own expenses and a generous grant from the Merton College Graduate Research Expenses fund.

Conflicts of Interest

The author hereby declares that they have no conflicts of interest, personal or professional, relating to this study.