Estimating the Micro-Targeting Effect: Evidence from a Survey Experiment During the 2020 U.S. Presidential Election

Musashi Harukawa

Abstract

The topic of micro-targeted political campaigning has been prominent in the public sphere since 2016. Despite a diverse literature warning about the dangers it poses to elections, democracy, and civic society, extant political science research indicates it may have no effect. Leveraging a novel design that incorporates both machine learning and causal inference, this paper estimates the effect of micro-targeted political advertising. This paper finds that among unaligned voters who had not cast their vote at the time of the survey, optimally allocating negative advertisements on the basis of respondent traits increases the proportion of respondents who dislike the target candidate by 8.7 percentage points, and decreases the proportion of respondents who intend to vote for the target candidate by 7.1 percentage points.

Introduction

Reports of the role of the data-driven campaigning firm *Cambridge Analytica* and Russian social media bots in the outcome of high-stakes political events of 2016 have created a very real need to understand the implications of the use of highly granular individual data for targeted political campaigns (Simon 2019).

Scholars in law, ethics and political philosophy view the consequences of technologically-driven political campaigning as damaging and considerable (see e.g. Witzleb, Paterson, and Richardson 2019). The implications of campaigning with highly personalized advertisements shown selectively to groups in the population, known as micro-targeting, are wide ranging, from threatening civic discourse through the creation of informational "filter bubbles" (Burkell and Regan 2020) to undermining the very foundations of democracy by weakening individual autonomy (Sunstein 2015). These arguments and warnings often rest on the assumption that micro-targeted campaigning affects electoral outcomes; "micro-targeting of voters can pay very handsome electoral dividends for a relatively modest investment" (Krotoszynski 2020).

Existing political science literature on political campaigning casts doubt on this assertion. A decade of field experiments has shown that the effects of presidential advertisements are small (Broockman and Green 2014), decay rapidly (Gerber et al. 2011; Mitchell 2012), and that these small average effects are not masking large heterogeneity (Coppock, Hill, and Vavreck 2020). On the other hand, psychological models of persuasion suggest that campaigns that know characteristics of the listener can increase the effectiveness of individually targeted advertisements (Cialdini 2007; Madsen and Pilditch 2018). It is difficult to conclude from these results whether highly-targeted political advertising campaigns can sway

an election to the extent that the normative literature asserts.

There does not appear to be a straightforward answer to the question "does micro-targeting *work*?" because such a question is itself not straightforward to investigate. All agents involved–from political campaigns and consultants to social media firms such as Facebook–are are unlikely to share their data or algorithms (Baldwin-Philippi 2017; Edelson et al. 2019). As a result, scholars resort to relying on proxies such as daily advertising prices (Liberini et al. 2020).

This paper presents a novel two-stage survey experiment designed to directly estimate the effect of micro-targeting. Using five real anti-Joseph Biden advertisements fielded by the Donald Trump campaign in the final month of the 2020 United States (US) presidential election, the design leverages 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. The difference between the first (control) and second (treatment) stage respondents is a causal estimate of the effect of optimal advertisement allocation on the basis of individual traits, i.e. micro-targeting.

The results show that the optimized allocation of advertisements leads to a significant boost in target outcome 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 estimated effect sizes found in this paper are greater than the margins seen in many crucial districts of the recent US presidential election. Given the significant implications this carries for electoral outcomes, I call for a renewed urgency in the stricter regulation of these methods, and greater transparency around targeted advertising.

Related Literature

Tailoring and *targeting* are two key concepts that I use extensively in this article, and whose meaning are used variably within the literature; I begin by clarifying the distinction. The *tailoring* of 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). Though not explicitly defined in the literature, I use the prefix "micro-" to refer to either of these being done on the basis of individual-level data, usually obtained via social media or a data broker. This article analyzes the effect of the advertisement allocation strategies of data-driven political campaigns, or *political micro-targeting*.

Targeted messaging in political campaigning has been a topic of scholarly interest since at least 2005 (Kang 2005), and the literature has developed considerably over the past decade. Psychological models of persuasion (Petty and Cacioppo 1986; Cialdini 2007) provide a basis for understanding the mechanisms by which messages may influence individual beliefs, attitudes, opinions or choice. These models emphasize that the persuasiveness of a message depends not only on the content of the message, but the traits of the receiver, their social context, and their attitudes towards source of the message. More recently these models have been used both by political data analytics companies to develop psychological models of voters, or "profiles" (Baldwin-Philippi 2017; Simon 2019). This strategy has been employed by firms such as *Cambridge Analytica* to match voters to the messages that are the most likely to be persuasive to them (Madsen 2019).

These events have naturally led to the question of whether more fine-grained targeting on the basis of individual traits, gathered via social media and other privacy-invasive tools, has resulted in more effective campaigning. Answering such a question, however, has been challenged by difficulty gaining access to political campaigns' targeting strategies, the granular individual data purportedly used in these models, or even the advertisements shown on Facebook themselves (Edelson et al. 2019).

There are strong reasons to be skeptical of the claims of campaigners and their ability to manipulate the masses. 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. Indeed, a recent large field experiment ($N = 34,000$, 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). It is difficult to square these null findings with claims of mass manipulation.

Nevertheless, scholars have found a variety of ways to overcome the difficulties of studying micro-targeting. Madsen and Pilditch (2018) leverage agent-based modeling to simulate a targeted campaign. Modeling voters as agents with Bayesian preferences and campaigns as a repeated game, Madsen and Pilditch (2018) shows that campaigns with knowledge of voter characteristics are more likely to succeed, providing evidence that targeting *should* work.

Other scholars leverage causal designs to measure the effect of advertising in counties at media market boundaries (Sides, Vavreck, and Warshaw 2020), finding an increase in advertising volume is associated with a small increase in vote share. Liberini et al. (2020) use daily advertisement prices on Facebook as a proxy for intensity of targeted advertising in the 2016 presidential election and find that exposure to

advertisements made individuals less likely to change their initial voting intentions.

The downside to these approaches is a reliance on proxies for micro-targeted advertising. Therefore, a third approach tests the mechanisms underlying micro-targeted persuasion. Zarouali et al. (2020) utilize an experiment ($N = 158$) to demonstrate this potential. Participants were asked to use a fake social media site, where short texts that they wrote were used to label participants as either introverted or extroverted and then shown a suitable advertisement accordingly. They find that the effect of correctly matched advertisements is stronger.

The main shortcoming of the approach of Zarouali et al. (2020) is an unnecessary reliance on the theoretical relationship between personality types and optimal advertisements. In order to show that micro-targeting works, we only need to show that it is possible to improve campaign effectiveness by optimally allocating advertisements on the basis of individual traits. This does not require us to label the intermediary states (i.e. "profiles") and assign accordingly, as we can instead train an algorithm to directly optimize predicted effect.

In contrast to the above studies, this study overcomes these difficulties by answering the key question as directly as possible: *is it possible to improve the effectiveness of a campaign by optimally allocating advertisements on the basis of individual traits*? If the answer to this question is no, then there is no room for micro-targeting to work. If the answer is yes, then I have shown that political campaigns are capable of improving the effectiveness of their campaigns by leveraging personal data. The potential consequences for democratic politics, from undermining elections (Burkell and Regan 2020) to the incentives for the abuse of data by elected officials (Krotoszynski 2020), is considerable.

Experimental Design

This experiment was fielded in a seven hour window on 28 October 2020. Respondents, recruited via Prolific, were directed to a survey website. Filters were set to ensure that respondents were United States citizens residing in the United States, and of voting age.

This study uses the United States as the setting for the experiment for the prevalence and variety of micro-targeted political campaigning present in the United States (Baldwin-Philippi 2017). Similarly, using the presidential election increased confidence that highly tailored advertisements would exist and be relevant to respondents.

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.

Respondents were then given a prompt stating that they would be shown a short advertisement. On the following page, respondents were shown one of five anti-Biden attack advertisements fielded by the Trump campaign in the month of October. These were chosen out of the set of all advertisements fielded by the Trump campaign between May and October, on the criteria that they were current and clearly tailored to different audiences. Details of the advertisements are in Table 1. Anti-Biden advertisements were chosen in favor of anti-Trump advertisements because of uncertainty surrounding Trump's health two weeks prior to running the experiment.

Table 1: Advertisements and Descriptions

Title	Description
"They Mock Us"	Clinton and Biden mocking of Trump supporters.
"Why did Biden let him do it?"	Advertisement about Hunter Biden's ostensible corruption.
"Biden will come for your guns"	2nd Amendment advertisement, stating Biden will take away guns.
"Insult"	Advertisement focusing on Biden's statement that Black Trump supporters are not Black.
"Real Leadership"	Advertisement focusing on wars and neglected veterans under Obama/Biden.

On the final page, respondents were asked for their opinion of Trump and Biden, and whom they intended to vote for. Given high rates of early voting, respondents also had the option of stating whom they had already voted for.

In the first stage, 1,500 respondents were randomly assigned an advertisement using permuted block randomization. The data gathered in the first stage was used to train five different models for predicting the most effective advertisement. The candidate models included Random Forest (RF), AdaBoost, Gradient Boosted Decision Trees (GBDT), Multi-Layer Perceptron Regressor (MLPR), and Support Vector Machine (SVM). These models were chosen in favor of generalized linear models either for their ability to learn highly conditional response surfaces, or a very fast fitting and predicting time.

The five models were tested using 30-fold cross-validation, and compared on their root mean squared error (RMSE), maximum error, and prediction time[1]. RF and AdaBoost outperformed the latter three algorithms on all of these metrics. Between RF and AdaBoost, RF performed weakly better and was less likely to predict ties, and was thus preferable to AdaBoost for predicting optimal advertisements[2]. This pre-fitted RF model was then uploaded to the web server. The transition from stage 1 to stage 2 was done in under an hour.

In the second stage, respondents answered the same pre-treatment questions. These answers were sent

[1]Prediction time was considered because the allocation engine could only handle single tasks, and the potential for long queues during traffic bursts was undesirable.

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

to a server-side python kernel, which given these values, used the uploaded RF model to predict Biden favorability for each of the five advertisements. The advertisement with the lowest favorability was then sent back to be shown to the respondent. The first and second stage are identical in all respects except the allocation mechanism for advertisements.

Hypotheses

The above experimental design can be described as follows. For individual i, we have pre-treatment covariates $\mathbf{X}_i$, post-treatment outcome Y_i, and they are shown advertisement $d_{i,a}$, where $a \in \{1, 5\}$. For convenience I will omit the subscript i from $d_{i,a}$.

In the first stage, advertisement allocation is randomized, with all advertisements equally likely: $a \sim \mathcal{U}\{1, 5\}$. The results of the first stage are used to fit the following predictive model:

$$Y_i = f(X_i, d_a)$$

As noted, five candidate algorithms are used to learn $f(\cdot)$: RF, Adaboost, GBDT, MLPR and SVM. The best-performing algorithm is used in the second stage to choose the optimal advertisement for any given vector of pre-treatment covariates X. Thus at the second stage, for individual i, we choose the advertisement $d^*_{i,a}$, that optimizes the predicted value of Y_i.

$$d^*(X_i) : opt_a \ f(X_i, d_{i,a})$$

Where in this case, $d^*(X_i)$ is the value of $d_a \in \{d_1, d_2, ...d_5\}$ that *minimizes* the predicted outcome $\hat{Y_i}$, given pre-treatment covariates X_i. Note that if this advertisements were promotional, then $d^*(X_i)$ would be the value that *maximizes* predicted favorability.

To estimate the effect of targeting, we compare the average outcome between randomly assigned stage 1, $\mathbb{E}_a[\mathbb{E}_i[Y_i(d_{i,a})]]$, and optimally assigned stage 2, $\mathbb{E}_i[Y_i(d^*(X_i))]$.[3] The difference between these two values is the *average treatment effect* (ATE) of targeting.

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

This design is used to test the effect of targeting on three outcomes that a campaign is likely to want

[3] *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.

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, favorability of candidate c is denoted $y_{i,c} \in \{1, 2, 3, 4, 5\}$, intent to vote for candidate c is denoted $v_{i,c}$, 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*): $\mathbb{E}_i[y_{i,Biden}(d^*)] < \mathbb{E}_a[\mathbb{E}_i[y_{i,Biden}(d_{i,a})]]$
- `Hypothesis 2` (*Micro-targeting Affects Voting Preference*): $\mathbb{E}_i[v_{i,Biden}(d^*)] < \mathbb{E}_a[\mathbb{E}_i[v_{i,Biden}(d_{i,a})]]$
- `Hypothesis 3` (*Micro-targeting Affects Turnout*): $\mathbb{E}_i[u_i(d^*)] < \mathbb{E}_a[\mathbb{E}_i[u_i(d_{i,a})]]$

These hypotheses are tested with the experimental design described above, where all participants are shown one of several advertisements from the same campaign, and the treatment is whether the respondent was targeted. In other words, individuals in the control group receive an advertisement at random, and individuals in the treatment group receive the advertisement predicted to have the largest effect on them.

For all of the above hypotheses, I account for the conditional effect of two pre-treatment covariates: partisan identification and early voting. 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, the effect of micro-targeting on undecided voters is a question of substantive and practical interest. Finally, the latter two outcomes are not measurable for individuals who have already voted.

Results

After filtering for irregularities and failed manipulation checks[4], the experiment provided a total of 2261 valid responses, with 1416 in the control group and 845 in the treatment group. A chi-squared test of independence found no dependence between the treatment indicator (stage) and any of the pre-treatment covariates, after conducting Holm (1979) and Benjamini-Hochberg (1995) multiple comparisons corrections.

The results are presented in three parts. The first part examines the predictive model fitted in the first stage. The second part looks at the effect of targeting on favorability for the complete sample. The third part looks at the effect of targeting on favorability, voting preference and turnout intention for the subset of respondents who had not cast a vote at the time of the survey.

[4]Irregularities consisted For complete descriptions of manipulation checks and balance checks, see Appendix III.

Stage 1 Results

Figure 1 shows the per-feature normalized Mean Gini Decrease[5] (nMGD) statistic given by the RF model used to assign allocations in the second stage. Interpreting this gives some insight into the relative importance of each pre-treatment covariate when predicting post-treatment Biden favorability. Although these statistics cannot be interpreted in the same way as regression coefficients[6], they do tell us the degree to which partitions on a feature produces 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.

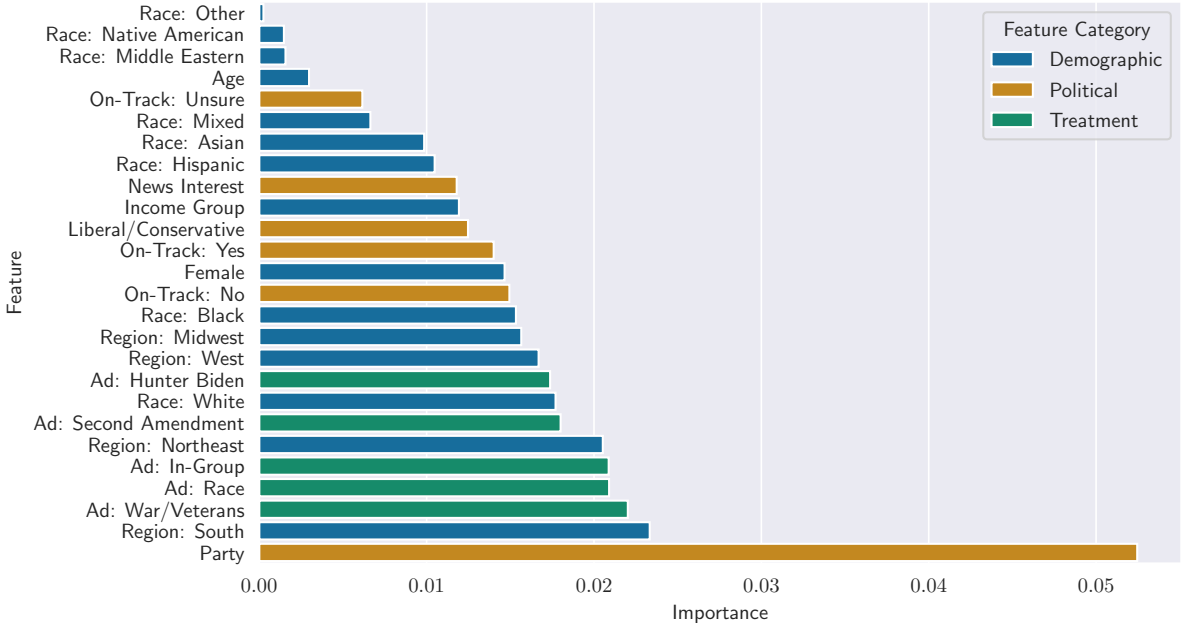

Figure 1: Stage 1 Normalized Feature Importances

Unsurprisingly, partisan self-identification was the most important feature in predicting average Biden favorability. However, the relatively high nMGD of the advertisement assignments presents evidence in favor of there being a meaningful effect of the advertisements. Given that these assignments were randomized, we can rule out both reverse causation and omitted variable bias as explanations for the interdependence of advertisement assignment and Biden favorability.

Stage 2 Results: Full Sample

I discuss only Biden favorability for the full sample, as intent to vote is not measurable for respondents who had already voted. Table 2 presents four linear models regressing the treatment, targeting, on Biden favorability, with the latter two columns interacting the treatment on partisan self-identification. The outcome is operationalized both as a five-point linear scale and a binary outcome indicating the

[5] As MGD is biased towards features with high cardinality, I normalize the feature importances by their cardinality.
[6] A causal interpretation of Stage 1 Results is discussed in the Appendix I.

proportion of respondents who dislike Biden.

	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

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

Table 2: Effect of Micro-Targeting on Candidate Favorability Interacted on Partisan-Self Identifaction

These models show an inconclusive set of results. The coefficient of interest, *Targeted (Unaligned)*, measures the treatment effect of targeting. In models (3) and (4), this is the effect of targeting among unaligned voters. The uninteracted model shows weak and insignificant results. The interacted model shows weak effects in the expected direction—a −0.11 decrease on the five-point linear scale (column 3) and a 5 percentage point increase in the proportion of respondents who dislike Biden (column 4). The significance on the final result, however, is not robust to multiple comparisons adjustments or the use of non-linear models (for tables with ordered logistic and logistic regressions, see Appendix II).

Though the coefficients on Democrat and Republican are large, significant and in the expected direction (3 and 4), the coefficient on targeting is roughly double for unaligned voters compared to the uninteracted effect, indicating that the uninteracted models (1 and 2) may be disguising some heterogeneity. This provides evidence to support the hypothesis that party identification and the opinions the respondent held on Biden prior to viewing the advertisement condition the effect of the advertisement.

Similarly, individuals who have already cast their vote for either candidate are less likely to form new opinions about Biden from a campaign advertisement. From the perspective of campaigns, their interest is to find the subset of individuals whose vote is changeable, and sway them; partisan individuals who have already voted are not that.

Stage 2 Results: Respondents who had not Voted

This section discusses the effect of targeting conditional on partisanship for the subset of respondents who had not already voted at the time of the survey ($N = 1,160$) for the three outcomes described in the section *Hypotheses*. Multiple models were fitted for each outcome: for Biden favorability, a five-point

outcome, a linear regression, ordered logistic regression, linear regression on a binary outcome, and logistic regression. For intent to vote for Biden and turnout intention, which were binary outcomes, a linear regression and logistic regression were fitted. The key results of these models are presented in Figure 2 (full results in Appendix II).

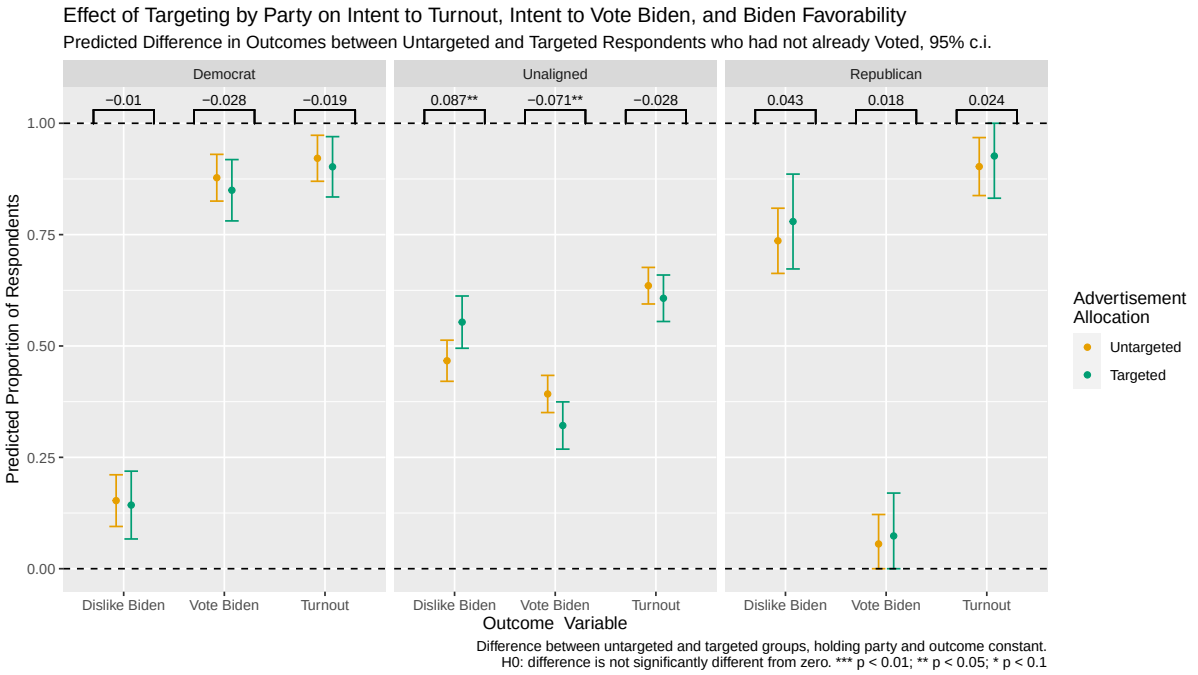

Figure 2: Effect of Targeting on Respondents who had not Voted

Figure 2 compares the predicted outcomes for the untargeted (control) and targeted (treatment) groups for each of the outcome variables, conditional on partisan self-identification. The value above each pair of points shows the difference between the untargeted and targeted outcome, with the stars indicating the p-value on a null hypothesis that the difference is not significantly different from zero. Note that these values are equivalent to the treatment coefficient in a regression of the outcome on treatment interacted on party, with the party as the reference category.

The key result is the first two pairs in the middle panel, "Unaligned". The pair on the left, above "Dislike Biden", shows that targeting causes an 8.7 percentage point increase in the proportion of respondents saying they dislike Biden ($SE = 0.038$, $p = 0.0228$)[7], whereas the pair in the middle, above "Vote Biden" shows that targeting causes a 7.1 percentage point decrease in the proportion of respondents stating they will vote for Biden ($SE = 0.034$, $p = 0.0397$)[8]. These results are fully robust to alternative operationalization of the outcome (see Appendix II).

[7]This result is robust to the inclusion of pre-treatment covariate ($\hat{\beta} = 0.066$, $SE = 0.033$) and remains significant at $\alpha = 0.05$ when accounting for multiple comparisons with the Benjamini-Hochberg correction (adj. $p = 0.0342$). Using the Holm correction, the p-value remains satisfactorily small $p = 0.0684$, although no longer significant at $\alpha = 0.05$.

[8]This result is less robust to the multiple comparisons correction ($p_{Holm} = 0.119$, $p_{BH} = 0.0595$), but I argue that it is not necessary to make this correction given that only a small number of pre-specified interactions has been tested for, and there is only one coefficient of interest (targeting on unaligned voters).

The remaining differences are small and insignificant ($p > 0.1$). For all groups, targeting appears to have little effect on overall intent to vote (Turnout). Importantly, identifying as either Republican or Democrat appears to nullify the effect of targeting on candidate preference and intent to vote for Biden. This may reflect that Democrats are less likely to be persuaded by anti-Biden advertisements, and Republicans already reaching maximal Biden antipathy.

The predicted proportions for all groups and outcomes are as expected. Democrats demonstrate the lowest predicted proportion of individuals who dislike Biden and the highest proportion of individuals who intend to vote for Biden, whereas Republicans reverse this pattern. Democrats and Republicans both have a high stated intention to vote (>90%) whereas for unaligned voters this is considerably lower (~60%).

Discussion

This experiment shows that micro-targeted campaigning can have significant effects on a crucial subgroup within the population: unaligned voters who have not cast their vote yet. Specifically, this paper finds that among voters with no partisan affiliation who had not voted yet, micro-targeting results in an increase of 8.7 percentage points stating that they dislike Biden, and a decrease of 7.1 percentage points stating that they intend to vote for Biden. The recent election shows that shifts of this magnitude can be extremely consequential in swing districts. I first address possible objections to the validity or significance of these results.

The significance of these results is largely robust to a variety of models and operationalizations of the outcome (linear, ordered logistic, or logistic) and the Holm and Benjamini-Hochberg corrections for multiple comparisons (Holm 1979; Benjamini and Hochberg 1995). Moreover, I argue that such corrections may not even be necessary, as the additional comparisons are being made for theoretically and substantively relevant reasons and the number of comparisons is relatively few.

A design-based objection to these results could be that unlike Zarouali et al. (2020) or political data analytics firms like *Cambridge Analytica*, the approach presented here does not attempt to infer psychological profiles or any other labels from the data. This is an intentional trade-off: by training a model to directly optimize the outcome of interest, I avoid relying on the validity of theoretical models and constructs, but in turn lose some explanatory power of *why* the targeting works. Given that my intention is to provide evidence that targeted political advertising *can* work, this approach is sufficient for demonstrating this potential.

Finally, this experiment does not prove that micro-targeting works in every context. This experiment looked at the effect of optimized allocation of five negative advertisements in a highly polarized campaign.

In contrast, Coppock, Hill, and Vavreck (2020) test 49 distinct advertisements and find that most advertisements are roughly equally effective, and find little evidence of heterogeneous effects. However, these results are not at odds; the only respondent characteristic Coppock, Hill, and Vavreck (2020) test for heterogeneity on is partisanship. The design presented in this paper additionally leverages heterogeneity in age, gender, race, income, region, ideological preferences and opinions on the state of the nation to search for optimal treatments. Moreover, given that Coppock, Hill, and Vavreck (2020) find all advertisements roughly equally effective on average, regardless of whether they are promotional or negative, I expect that these results can be replicated using a different set of advertisements.

Having addressed these limitations, I believe that this experiment satisfactorily shows that micro-targeting can have substantial and consequential effects on the preferences of swing voters in an election. That this effect is found among unaligned voters who had not voted yet agrees with the existing literature on persuasion; a desire for consistency restricts the effect on Democratic voters (Cialdini 2007), and may not cause any kind of attitude or opinion shift for Republican voters because it simply reinforces their priors (Madsen 2019).

The magnitude of effect seen here is wider than the margin seen in several swing states in the 2020 election, indicating that targeting could have very real impacts on electoral outcomes. A follow-up experiment is needed to test additional hypotheses: is this effect particular to the tumultuous and charged climate of the 2020 US presidential election or its two-party system? Would greater disclosure of the fact that the advertisement is targeted nullify the effect? These questions are essential to discovering effective methods for regulating this technology and curbing the impact it could have on democratic processes.

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). This paper 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 of the first stage 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 will instead be tested in a follow-up experiment.

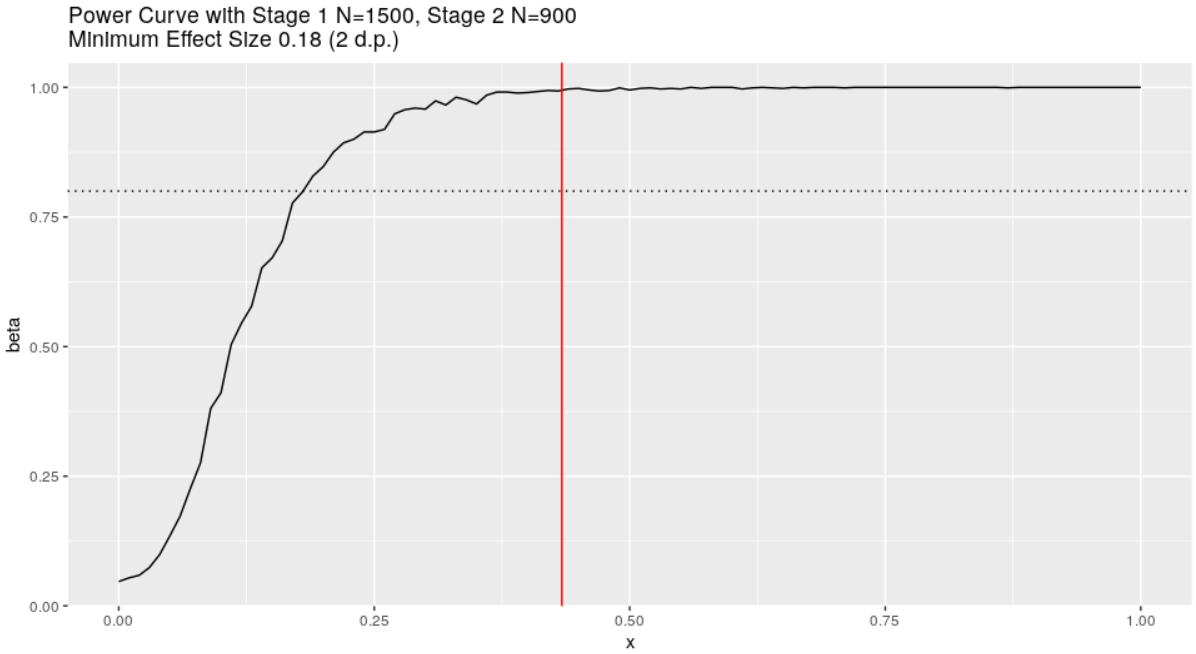

Figure 3: Power Analysis from CHV20 Data

As noted in the main body of this article, 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?

The first difference to note is that whereas CHV20 focuses on between-voter heterogeneity, my experiment focuses on within-voter heterogeneity. Put differently, even if the effect of an advertisement does not vary greatly from person to person, it is still possible for the effect of different advertisements to vary for the same person.

The targeting algorithm in this paper leverages this within-voter heterogeneity. If we represent the outcome for voter i watching advertisement d as $Y_i(d)$, then the targeting algorithm is calculating $Y_i(1), Y_i(2), ..., Y_i(d)$ and realizing the value of d that optimizes Y_i.

However, if all advertisements do not vary at all for all individuals, then all values of $Y_i(d)$ would be fixed for all i and d. Given that my experiment finds variance over d, this is apparently not the case. The explanation for this difference, as I state above, is that CHV20 do not test for heterogeneity on respondent characteristics other than partisanship.

Causal Interpretation for Stage 1

Randomized advertisement assignment in the first stage allows for a causal interpretation of the coefficient on each advertisement, but it is unclear which advertisement should serve as a reference category. 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: Full Regression Results

The following appendix contains the full results of the various models and operationalizations.

Effect on Candidate Favorability

Table 3 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 $\alpha = 0.05$ ($p = 0.0192,\ 0.0338,\ 0.0228,\ 0.0416$).

	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

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

Table 3: Effect of Micro-Targeting on Candidate Favorability, Interacted on Partisan Self-Identification among Respondents who had not Voted

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. Table 4 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.

	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

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

Table 4: Effect of Micro-Targeting on Intention to Vote for Biden among Respondents who had not Voted

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.

	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

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

Table 5: Effect of Micro-Targeting on Turnout among Respondents who had not Voted

Effect on Turnout

The final set of results, shown in Table 5, 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.

Appendix III: Implementation Notes

Randomization Check

The causal interpretation of these results rely on there not being any systematic differences between the treatment (targeted) and control (untargeted) groups. That the control data must be gathered to run 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. Table 6 shows a Chi-Squared test of independence on assignment to treatment or control against all of the pre-treatment covariates. All hypotheses fail to achieve significance except for ideology, but this is not robust to the Holm (1979) or Benjamini-Hochberg (1995) multiple comparisons 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 6: 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

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. Figure 2 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.

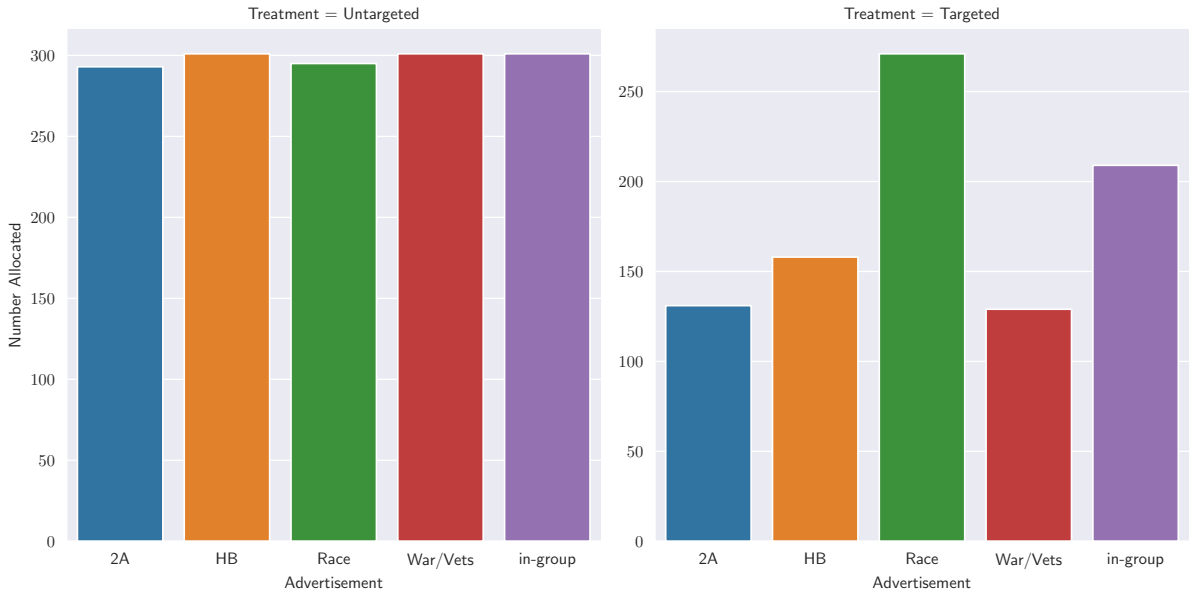

Figure 4: Advertisement Allocations

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[9] 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 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, this 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`.

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

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 US 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 PhD 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.

Bibliography

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. 2020. "Voting Public: Leveraging Personal Information to Construct Voter Preference." In *Big Data, Political Campaigning and the Law: Democracy and Privacty 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).

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

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.

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 Privacty in the Age of Micro-Targeting*, 186–213. Routledge.

Liberini, Federica, Antonio Russo, Ángel Cuevas, Ruben Cuevas, and others. 2020. "Politics in the Facebook Era-Evidence from the 2016 Us Presidential Elections."

Madsen, Jens Koed. 2019. *The Psychology of Micro-Targeted Election Campaigns*. Basingstoke.

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.

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.

Sides, John, Lynn Vavreck, and Christopher Warshaw. 2020. "The Effect of Television Advertising in United States Elections."

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."

Witzleb, Normann, Moira Paterson, and Janice Richardson. 2019. *Big Data, Political Campaigning and the Law: Democracy and Privacy in the Age of Micro-Targeting*. Routledge.

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*.