Special thanks for this thesis are due to many more people than I can feasibly include. However, I would like to thank Tobias Winemann for thoughtful advice in topic selection; Felix Iglhaut for helping me through initial conceptualization of this paper; Douglas Ahler for providing me with data necessary to drive results; and Matthew Pigott, Mark Barrett, Paul Wallace, Ryan Sanders, and Shaunt Gharabegian for help with editing. Thanks is also due to my parents, whose endless love and support have made it possible for me to pursue a love of economics far beyond what I thought would have been possible for myself. I would also like to thank the broader Bocconi community for its assistance and resources to help me finish this project, and, most of all for this project, Nicola Gennaioli for excellent advisorship.

1 Introduction

Despite the rancorous divisions which characterize contemporary American politics, there is widespread agreement among members of the public that political polarization is one of the greatest present threats to democracy. Members of both parties often describe issues related to differences in opinion as evidence of ever-growing Balkanization that, if allowed to continue unchecked, will result in nothing less than a bifurcation of the country along a red-blue divide. Unfortunately, in referring to political polarization, it is also often the case that members of the general public use fuzzy and unhelpful definitions, which often amount to saying that the type of polarization they dislike is the kind that mobilizes members of the other party against their own. The logic of polarization seems to convince partisans of all persuasions that there is a genuine and widening difference between the centers of both parties and simultaneously, that their party is not the one which deserves most, or perhaps any, blame. The sentiment that partisanship is getting more vicious with each passing day has inspired many works of political science and economics[1] which attempt to properly diagnose this malaise at the heart of the American two-party system, but has seldom yielded clarity around the phenomenon.

Understanding the factors behind political polarization is as difficult as understanding the universe of reasons which motivate human decision making processes in any setting. Political polarization evades careful scrutiny not simply because the definitions can be opaque, but also because the theorized impact of political polarization is contingent upon its definition, which causes estimates to range from profound to nonexistent. Debates between political scientists like Fiorina and Abramowitz in the late aughts demonstrate the sharp disagreement at play when examining political polarization. While Fiorina points out the high degree of similarity present among members of the American polity and consistency of beliefs over time as evidence of a lack of polarization (Fiorina et al. 2011), Abramowitz and co-authors provide evidence that, among issues which have been consistently measured since the 1960s, there has been a large degree of ideological separation (Abramowitz et al. 2008) over time. These two are far from the only members of the scientific community to have such disputes, or to characterize polarization in different ways.

Explanations for political polarization have emerged from a host of different social sciences. Economic explanations might hold that anxiety around household well-being explains more recent moves in polarization, such as the election of Donald Trump in 2016. Psychological explanations would posit that differences in “Big 5” personality traits drive party selection, and that these contrasts underlie the political differences that we observe. This thesis will take a different tact, however. I posit that political polarization is driven by heuristic thinking, i.e., a tendency to use mental shortcuts and quick decision rules which frequently fail to appreciate the nuance of a situation and, thus, give an incorrect answer. Although I believe the differences in heuristic thinking would extend to many well-known cognitive biases, my focus here is on stereotyping. Thus, heuristics in the sense of this paper suggest an unawareness of the other party, or an unwillingness to see it as it truly is. Following the reasoning of Gennaioli et al. 2019, I examine the underlying connection between polarization and heuristic thinking by attempting to prove that social groups which exhibit higher levels of polarization also stereotype more. To uncover individuals’ propensities to become polarized, I use the introduction of the internet as (possibly heterogeneous) treatment effect and provide

[1] Including this one

estimates for the changes in affective polarization which occurred between 1996 and 2012. To figure out how to properly account for treatment heterogeneity, and to do so in a way which does not require snooping or multiple hypothesis testing, I employ new methods in the causal forest literature from Athey, Tibshirani, et al. Finally, I apply those groups which demonstrate higher levels of polarization to a study on the strength of out-party stereotyping from Ahler et al. 2018. It is my belief that there will be significant differences in polarization among members of the population (in particular I expect to find that there are differences along respondents' ages), and that these differences will also map on to an increased propensity to stereotype. Evidence in favor of this theory would also imply that heuristically-motivated thinking is a strong driver of political polarization.

There has been much discussion in the literature as to whether or not the internet has been a polarizing or depolarizing force. This paper shows that, if the introduction of the internet into homes is regarded as a treatment which obeys conditional unconfoundedness, then there is reason to believe that the internet did increase the polarization among Republicans and Democrats, and that the effects of the internet were not symmetrical. It is necessary to clarify what is meant by political polarization, at least in the context of this paper. In truth, every definition will be wanting for some clarity, or ignore some case which seems important. Rather than attempt to synthesize a broad body of literature into a single phrase, I will limit my focus onto one immediate and well-understood dimension of the literature: affective polarization, or the difference in the emotional attachment that individuals feel towards their party compared to the other. Affective polarization can be measured by looking at the affective gap in survey questions from the American National Election Survey. Even after defining a limited problem for which tractable answers exist, there is still the lingering issue of developing a model which can account for the growth of political polarization, as has often been done in the social science tradition. However, I believe that the determinants of polarization are a mess of complicated dependencies between policy issues, identities (Mason 2013), and the ways in which those identities interact with the outside world. Instead of providing a model whose veracity will be proved or disproved by the procession of this thesis, I employ recent developments in machine learning literature which are uniquely well-suited to the problem of detecting heterogeneity, namely causal forests (Athey, Tibshirani, et al.), to "model" polarization nonparametrically, attempting to understand its determinants through the rich body of survey data which is already available, rather than through a presupposed model[2]. Through such methods, polarization is more properly viewed as the complicated and messy process that it is, and several surprising variables emerge as being important to the process. The remainder of the thesis proceeds as follows: section 2 covers the history of the literature synthesized in this work, helping to explain the factors which contribute to the social-identity conceptualization of polarization and its grounding here; section 3 describes the sample population from ANES used in this paper, and covers the methodology of causal forests which this work employs; section 4 summarizes my analysis of the ANES and Ahler et al. 2018 data; and section 5 concludes with discussion on my findings.

[2] In general, I attempt to limit the use of the word "model" as much as possible within this work for precisely this reason. The methodology of this paper uses algorithmic ensemble learning, thus abstracting away from the usual functional trappings of a model, e.g., linearity.

2 Literature Review

There are, broadly, three strands of related literature to this thesis: work on stereotypes and social identity in political economics, studies on affective polarization, and machine learning methods. The basis for the use of stereotypes in political decision making has its roots in Social Identity Theory (SIT) as first proposed by Tajfel et al. 1974. Under this model, individuals possess strong preferences for the well-being of in-group members over and against the interest of out-group members. By its nature, this notion extends to the political realm rather naturally. Shayo 2009 introduced the concept of a social identity equilibrium in a variant on the standard Melzer-Richard redistribution model wherein agents possessed social identity preferences as well as redistributive ones. Within that framework, agents' bliss points were dependent upon the social group that they were a part of, both in the sense that being a member of one group meant that individual had different preferences than they would have were they in another, and in the sense that the relative status of the group was a salient factor in determining the strength of association. Further empirical work by Shayo and others has demonstrated that introducing social identities into political economics provides insight into otherwise puzzling behavior[3].

This thesis draws most of the theoretical framework from Bordalo et al. 2016 and Gennaioli et al. 2019, which lay out the foundation for the stereotype-based conceptualization of political distortions. Stereotypes in these works are distortions in the true likelihood ratio for certain traits which are inflated by representativeness, and modeled as

$$f^{\theta}(X|G) = f(X|G)\left[\frac{f(X|G)}{f(X|\bar{G})}\right]^{\theta} Z$$

where G describes the group that the individual is considering, $\bar{G}$ represents the group used for contrast, and $\theta (> 0)$ is a parameter which expresses the intensity of stereotyping. Thus, agents in this framework tend to believe that characteristics which appear relatively frequently in a population subgroup are much more likely within that subgroup than they actually are. The typical example of this phenomenon is the tendency to believe that a stereotypical Irish person has red hair. In truth, only about 10 percent of Irish people have red hair, but because only around 2 percent of the general population has red hair, orange locks become representative of Ireland. Stereotypes of this nature have the so-called "kernel of truth" property, which expresses the idea that distorted distributions highlight differences between groups in the direction of those differences. That is to say, while individuals' stereotyped notions of the differences between groups is, in general, correct about the direction of difference, it often overstates the contrast. In policy terms, the belief that inter-group differences are important and are more stark than they are in reality implies that agents may prefer policies which are detrimental to more intra-group members than expected, that agents may be too pessimistic about their own status within society when voting, and that the political class will be focused on issues which draw the largest contrast[4]. That work also grounds the observed extremity in voters' policy beliefs (as well as large shifts in preferences recently observed in the US and France) as being driven by individual social identity and, more particularly, with the sudden shift in the

[3] See Atkin et al. 2019 for an example of this - demonstrating how the consumption preferences of Hindus and Muslims in India change around instances of sectarian violence

[4] See Fiorina et al. 2011 for more on the argument about the political class, and Gennaioli et al. 2019 for more on the notion of large contrasts among groups and their importance to politics.

axis of political cleavages from economic to cultural. Such a viewpoint explains recent findings around the correlations between social identity and a host of beliefs (aligning with the work of Alesina et al. 2018 and Kahan 2015) This thesis cannot positively identify whether an individual voter chooses to identify with one group or another, but I attempt to mimic this analysis by dividing individuals according to the plethora of possible identities that ANES allows me to consider.

This thesis also draws from the preexisting literature on affective polarization, which dovetails with past work on social identity theory. Iyengar and Westwood 2015 define affective polarization as "the tendency of people identifying as Republicans or Democrats to view opposing partisans negatively and copartisans positively"[5]. Thus, affective polarization is a "natural offshoot of this sense of partisan group identity" (Iyengar, Lelkes, et al. 2019) and is captured by the contrast in sentiments (or feelings) which party members hold between their party and the opposing party. Evidence of this kind of partisan behavior is widespread, and not limited to the political realm. For example, politically homophilic behavior has been observed and demonstrated in online dating (Huber et al. 2017). However, while it is taken as given in contemporary American politics that the parties (and their constituent members) are more divided now than they have been since the Civil War, the evidence of this polarization is far more sanguine. Morris Fiorina argues in the 2005 book Culture War that the differences among party members are not nearly as extreme as many people suppose them to be, but that the political class itself may be significantly polarized (Fiorina et al. 2011) . Hence, the reality of polarization is driven by the supply of politicians, rather than the demand. This work motivates Iyengar and Westwood 2015 develop Implicit Association Tests for partisan bias, which would capture the extent of partisan affect while disentangling political polarization in voters from political polarization in politicians. More recently, Boxell et al. 2017 examines the role that the internet has played in driving polarization. The paper compares the distribution of internet usage by age groups to the distribution of polarization (according to their measures) by age, and ultimately finds that the respondents who were the most likely to be polarized according to their measures, i.e., the elderly, were also the least likely to use the internet. From this evidence, the paper concludes that if a causal link exists between the internet and polarization, its channel is more complicated than has been appreciated in contemporary discussion. In addition, Boxell et al. 2019 finds that the United States is a particularly polarized country among nine OECD countries studied, which weakens the argument that the internet would be a particular catalyst of political polarization since the internet is ubiquitous. If anything, according to that paper, recent increases in polarization in the U.S. can be more plausibly traced to the founding of Fox News in 1996. This thesis will explore the question of internet-related polarization through the methods of causal forests, which can flexibly account for non-obvious causal pathways (even if those pathways remain unknown after the fact) in estimating treatment effects from internet usage on polarization.

A related set of work in affective polarization neatly combines the social identity (i.e., individual-within-a-group focused) approach with polarization due to stereotyping. Ahler et al. 2018 demonstrates that survey respondents hold considerably over-exaggerated views on party-prototypical groups, such as the incidence of Democrats who are members of the lesbian, gay, and bisexual community and Republicans who earn more than $250,000 a year. The paper demonstrates that these stereotypes are particular to the parties and, in

[5] this is the working definition which will be adopted throughout this work when mentioning affective polarization, or indeed polarization, unless otherwise specified

line with Gennaioli et al. 2019, these stereotypes exhibit a kernel of truth. Furthermore, the responses are not due to other artifacts in reporting that one may expect[6]. Work in this space is quite important because it demonstrates that individuals have distorted notions of the composition of opposing parties (and, indeed their own, though the effects are more modest), which I will argue is a source of polarization. Setting aside complications due to non-political attitudes (e.g. racism and sexism), one might suspect that a Democrat who believed that 60% of Republican party members earned more than $250,000 a year would hold more emotionally-charged disdain towards policies proposed by the Republican party which favored that group. While that paper demonstrates that there are important correlations between group membership and stereotypical thinking, this thesis builds upon that work by providing more evidence that groups with higher levels of stereotypical thinking overlap neatly onto politically polarized groups along certain dimensions.

The final strand of literature which bears mentioning in this section is that concerning machine learning; in particular, this paper draws from the literature on causal forests, a special case of random forests. The original idea for the random forest has its foundations in Breiman 2001, which builds off of classification and regression trees from Breiman et al. 1984. As suggested by the name, trees can be used both in classification and regression problems, but the focus for this paper is on regressions. Every tree begins with a subsample of the entire data set collected together. The random forest algorithm splits this first, largest "parent node" into two smaller "child nodes" based upon a splitting criteria[7]. From that point, each of the child nodes is considered to be a parent node, from which the algorithm finds another split, generating two more child nodes which offshoot from the original child (now considered "parent") nodes. This process continues recursively until the algorithm cannot find a good split, given the complexity penalty it faces (which punishes trees with too much depth). All nodes which do not end the tree are called "non-terminal", else they are "terminal". Fitted values for the observations are given by the fitted value for the terminal node that the observation falls into. As with classical regression, there are ways of quantifying the fit of the algorithm according to predictive errors, in this case taken at each terminal node. By their nature, regression trees tend to flexibly model observations in a way that classical (linear) regression does not. The major shortcomings of such trees is that the individual trees, though unbiased, have considerable variances and overfit the data. To advance this methodology forward, Breiman 2001 proposed random forests, which is a regression or classification algorithm that grows many trees (usually around 500) and which outputs fitted values according to the average across all trees.

Crucially for my purposes, modern machine learning algorithms attempt to incorporate elements of subgroup identification into the nonstandard regression literature. Su et al. 2009 proposes a method to flexibly identify population subgroups of interest by recursive partitioning. Dividing the population into subgroups according to the data, rather than by prior expectations of the researcher, is desirable for a few different reasons. First, it directs attention to the population groups which demonstrate the most pronounced treatment effects (or, in a non-treatment setting, the groups which have a high degree of difference with the others), which may help to understand the causal pathways of a given effect. More importantly, such methods create partitions without the need for

[6] Specifically, these distortions do not arise as a result of innumeracy, ignorance of base rates, or expressive responding

[7] In the traditional CART algorithm, the best split is the one which most reduces node impurity, usually quantified by the Gini impurity measure

researchers to snoop through data beforehand or apply harsh penalties to confidence intervals as a way of accounting for specification search or multiple testing. In essence, a researcher employing these methods has the ability to observe strong differences in effects without needing to preregister the particular subgroups which will demonstrate pronounced differences. Further efforts have been made at obtaining credible estimates for the Conditional Average Treatment Effect (CATE) of population subgroups within the literature. Chipman et al. 2010 proposes "Bayesian Additive Regression Trees", which use nonparametric bootstraps of the population to generate posterior draws for the CATE. Although this method does provide estimates for the treatment effects of interest, it has two major drawbacks. First, there is no guarantee that the estimates converge to the true CATE. Secondly, the method does not provide for any way by which to carry out traditional hypothesis testing (which is exacerbated by the lack of a guarantee of posterior convergence). To address these issues and recover more traditional hypothesis testing, Athey, Tibshirani, et al. builds on regression tree methods with "causal forests", which allow the user to obtain partitions of the data without snooping, and in such a way as to avoid problems of multiple hypothesis testing that would usually arise when pre-registration plans are not in place. Causal forests differ from regression forests due to the nature of the trees that each one grows. Random forests use standard CART methods, and thus are ensembles of "random trees". Causal forests are grown from causal trees, which are deliberately designed to detect treatment effect heterogeneity by splitting parent nodes according to heterogeneity in the outcome variable. The founding paper for the method also demonstrates that causal forests allow for conditional treatment effect estimation with proper standard errors, which are shown to be asymptotically unbiased for the true conditional distribution, as long as the trees grown are "honest", a topic which will be examined in depth in the next section.

3 Methodology

3.1 The American National Election Survey

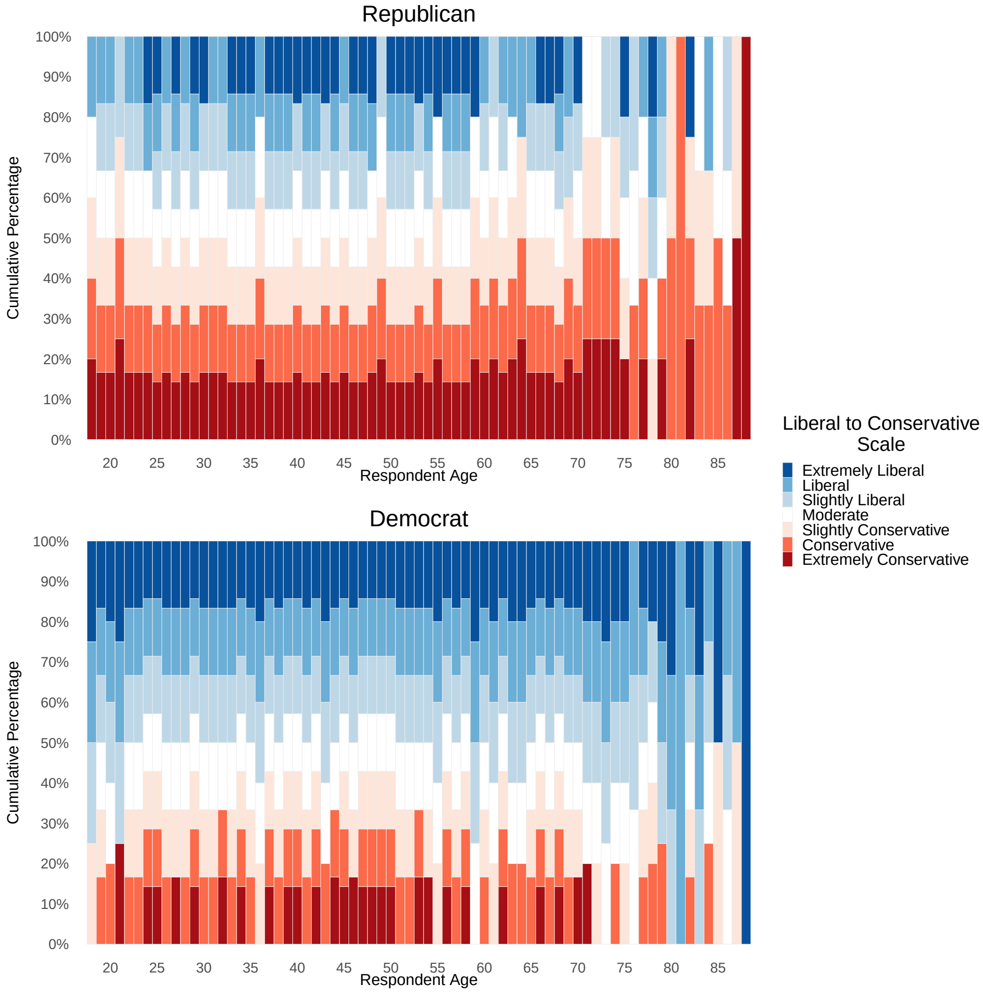
The main results of this paper all use data from the American National Election Survey (ANES) cumulative time series file. The time series covers data from ANES surveys ranging back to 1948, but this thesis only considers data between 1996 and 2012, when questions about household internet access were asked. Additionally, responses for the year 2002 are removed because the suite of questions asked sharply differs from the other waves of the study, as the 2002 wave of the study was administered only for the purposes of creating a panel study from 2000 to 2004. ANES is considered by many to be the gold standard of election-year voter surveys in the US, and its use here is motivated by the richness of the data set, its availability in every presidential election since the creation of the internet, as well as the possibility to generate the most direct comparison with the work of previous papers dealing with polarization, e.g. Abramowitz et al. 2008. To provide early demonstrations of the survey population, figure 1a shows answers one-to-seven scale of respondents' beliefs about the liberalism or conservatism of the major parties against the age of the respondent, and figure 1b shows violin plots of partisans' thermometer ratings, out of 100, for the other party over the time frame studied. These thermometer ratings bear some explanation, as they are quite important to my overall specification. Every thermometer rating asks respondents to place the subject of the question on a scale from 0 to 100, where 0 indicates the coldest feelings towards that subject, 50 indicates no feelings, and 100 indicates the warmest feelings. The violin plots establish the baseline trends in out-party affect during this period. Partisans have been souring on their political opponents over time, and the number of respondents who indicate the coldest possible feelings for the opposing party (i.e., a score of 0 on the thermometer rating) has been increasing in the past three decades.

In order to build a data set which could be used by the generalized random forest algorithm, several compromises needed to be made between maximizing observations and minimizing personal interference with the data which have bearing on the reliability of the estimates obtained and their interpretation. First, attention is restricted to only the variables which were asked in every wave of the study within the time interval of interest. As a consequence of this decision, questions which were phased out in the middle of this period or which were introduced during this period are ignored, which may have the effect of missing certain variables which were introduced to the study to respond to changing political circumstances, or to incorporate new knowledge about the voter base. For example, VCF9244 asks respondents about their willingness to trust others in society, and would be good inclusion to understand how these respondents interface with the general culture in line with Tabellini 2010. Unfortunately, this measure was only introduced in 2008 and cannot be used. Secondly, in some cases where a variable was administered to only a half-sample, but where the other half was given a series of branching questions which, taken together, result in a sensible answer to the original version of the question, the branching answers are coded back into responses for the non-branching, time-series version of the variable. For example, time series variable VCF0839 records the responses to a question where respondents place themselves on a one-to-seven scale expressing their opinion about government spending on social services. In 2000 some respondents were asked this same question, but others were asked two questions about the same topic; first, respondents were asked to sort themselves into one of three groups ("for", "against", and "neutral") and then to express the strength of their conviction on the matter ("strong", "moderate", "weak" in the case of being for or against in the first question, with no follow-up if the initial answer was neutral). Even though these responses map back into a one-to-seven scale in theory, recoding these answers according to the normal scale may introduce problems with construct validity, as it is impossible to know if a respondent would give the same response to the question whether presented a "branching" version or not. The normal way to correct a problem like this would be to cluster these observations together to account for the error introduced by recoding responses in this way. While I do cluster my observations at the yearly level, which helps to account for the error introduced by having a less reliable predictor for these individuals, sample size limitations make clustering at the level of "half-sample participation" unfeasible for estimation. That is, because each respondent was chosen randomly for half-sample participation for each question, there would need to be clusters in each year which cover every possible combination of half-sample participation. With an average of 343 observations in each year entered into the algorithm, these clusters would be quite small and the reduction in standard error bias would unfavorably trade off with the increase in estimated standard error variance.

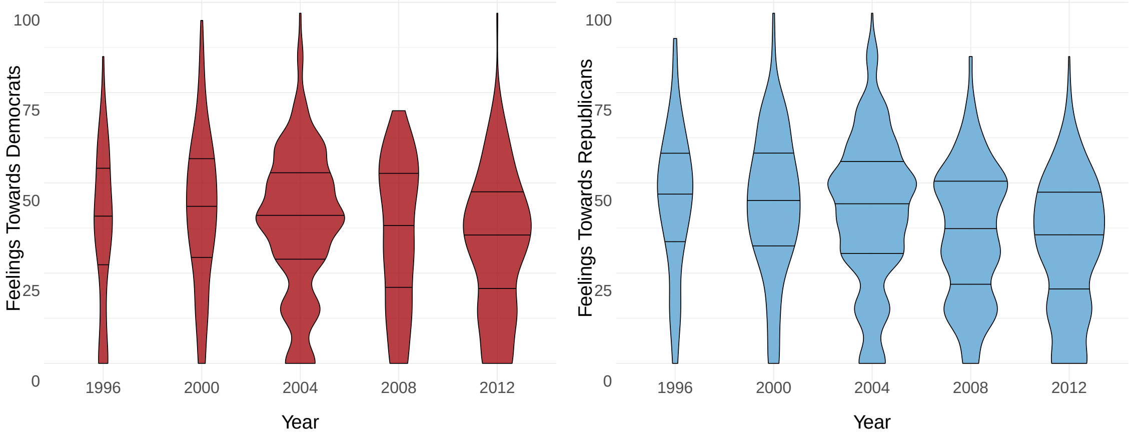
The reason behind such a low number of observations and the need to drop variables which are not common to every wave brings me to the final, and most troubling, concern. Generalized random forests do not have a way to deal with missing values in their current implementation, so all observations with a missing value for any predictor within the set of "common predictors" (i.e., the variables which were in every wave of the study over the

Figure 1: Two Plots Describing Trends in Party Sentiment

(a) (Weighted) responses to the seven point liberal-to-conservative scale for each of the major parties, by age



(b) (Weighted) violin plots of the sentiment towards the other party. The graph on the left shows results for Republicans, while the one on the right shows results for Democrats. The widths of the plots show the distribution of responses by year. Lines through the violin plots are, in ascending order, the 25th, 50th, and 75th percentile of the data for each year



time frame of interest) are removed. This causes issues with the reliability of estimation in two distinct ways. First, this restriction reduces observations within this set from around 5,000 to 1,677. Secondly, indiscriminately removing observations due to a missing value for any predictor likely introduces significant selection bias into the data. Other machine learning methods, such as CART (Breiman et al. 1984) deal with missing data through the use of imputation, and something like this is planned for the generalized random forest, but was not ready at the time of writing. In the future, such imputation could be used to obtain a larger sample size than is presented here and to help minimize (though not completely eliminate) issues related to selection bias within this population. The particular structure of ANES questions also creates a more subtle problem for the data when missing values are removed. For some variables, respondents who answer "don't know" to a question may be coded as providing an "NA" response to the variable, and would subsequently be dropped from the data before entering into the random forest. In cases where something like this occurs, careful attention has been paid to this problem and responses coded as "NA" enter into the data set as binary variables. However, this cannot be done for every variable of every data type as it runs the risk of flattening non-applicable responses of all types into definite responses of uncertainty. There are, thus, two possible interpretations for the direction of bias introduced, and one expectation to be verified. On the one hand, a sample population which is more politically informed than the general population is the most likely to be polarized (see Ahler et al. 2018), and the introduction of the internet would create the type of echo chambers Sunstein 2002 which further drive polarization and thus overstate the treatment effects. However, it could also be the case that individuals using heuristics to evaluate the parties do not respond to changes in their underlying level of information. In those cases, the introduction of the internet would likely have no effect because it would merely calcify pre-existing beliefs, but not change them. I would also expect that the algorithm will place a premium on information concerning the political knowledge of respondents. VCF0050a and VCF0050b report the level of political knowledge that the interviewer felt the respondent possessed in the pre- and post- election surveys respectively. Assuming for illustrative purposes that all respondents were rated in the top two categories, these variables would provide a binary distinction between relatively "high" and "low" information voters, which I expect to be an important difference in the data generating process and, ideally, the algorithm too. This in turn may cause the "low" information voters to have lower fitted values than is realistic, for "high" information voters to have systematically higher fitted values, or for some combination thereof. There are, naturally, caveats to this reasoning, beginning with the fact that our data set is not a clean binary between the two most informed voter types. Moreover, the fact that the level of political knowledge is subjective means that this variable could be relatively unreliable inasmuch as the personal unconscious biases of the interviewer may sway categorization of the respondent. On this point, more analysis would need to be done to examine the role that biases may play in the interviewer's evaluation. The measure is nonetheless included, but one must be attentive to the fact that "high information" within this set may not mean the same thing as high information in the population. As I later explore in analysis, the evaluated level of a respondent's political knowledge can help clarify other differences, but is not a large source of difference in and of itself.

The goal of estimating affective polarization is quite difficult in theory. It can be challenging to obtain a construct which is capable of expressing the difference between sorting, wherein partisans tend towards the party with which they are the most closely aligned, and polarization, wherein partisans express more pronounced negative feelings towards out-group members. Often, what looks like affective polarization is actually sorting in disguise; i.e., yearly trends which demonstrate a lower affect towards the out-group might be the result of partisan sorting into the correct parties, a trend which has been well-documented in American politics at least going back to the book The Big Sort . To avoid issues like this, Boxell et al. 2017 designs a suite of measures of polarization and sorting which are combined into an overall polarization index for a given year. I would argue that these measures, which aggregate for all parties, may fail to capture polarizing behavior at the extremes. This is because the index's component measures, such as the "partisan affect polarization" frequently estimate means of the population[8], while polarization may be thought of as a widening in the tails of the distribution. That is to say, it is not necessarily the case that political opinions need to be truly bimodal in order for us to conclude that polarization is a present force, as is sometimes conceptualized. There does not need to be "Two Americas" for there to be a clearly polarized One. This is for two reasons: first, the every day use of the term "polarization" refers to the widening gap between the parties and is subject to perception of those parties' rhetoric, which is often enforced by the behavior at the extremes of the distribution in light of research demonstrating large ideological polarization among political elites[9]. More importantly, the type of divergent behavior which results in a bimodal distribution may at first look like movement in the tails, particularly in light of evidence that the most polarized members of the parties are the elites who drive further polarization (Fleisher et al. 2004 among others provides evidence of this kind of polarization). It would be erroneous to conclude that polarization is not a strong force simply because we do not observe bimodality. In some ways, we ought to conceptualize polarization as a process whose completed outcome is perhaps distinct bimodality, but which is moved first from the tails.

3.2 Causal Forests

To tackle this problem, the main results are all reported using the method of causal forests first covered in Athey, Tibshirani, et al. As with all random forests, this is an ensemble learning method which allows for the growth of many "trees" (i.e., partitions of a random sample of the data) that have low bias and, through the collection of many estimates into a forest, low variance on the treatment effect estimates. All of these methods grow trees on a sampled subset of data, and thus every observation in the sample is present in some trees and "out of bag" in others. This methodology was chosen for a few different reasons based upon the research question. First, the use of any recursive partitioning method allows researchers to identify population subgroups of interest without the need to specify or preregister those groups themselves (Su et al. 2009). Such identification is highly desirable in a setting such as this, because the determinants of political identity are complicated and unlikely to stem from immediate sources. Second, as a machine learning method, causal forests allow for the possibility of estimating polarization within groups in a nonparametric fashion, which is chosen in this setting precisely because the question of interest is on whether or not important and different groups exist. By contrast, a modeling approach would only allow me to identify whether groups of my

[8] In fact, partisan affect polarization is, by definition, a weighted sum of the difference in thermometer ratings between the parties which, under an assumption of normality, would be the unbiased estimator for the population mean

[9] This is particularly true if one believes that perceptions are distorted by more representative tails as in Gennaioli et al. 2019.

own aggregation are different in a meaningful way, which suffers both from being a less interesting question, as well as being subject to personal subconscious bias. Random forests allow me to overcome both of those shortcomings. Finally, causal forests are chosen over related methods (such as Bayesian Additive Regression Trees, see Chipman et al. 2010) because causal forests allow for the possibility of hypothesis testing in a setting which allows for clustering, and the method provides standard errors which are asymptotically valid for such testing (Athey, Tibshirani, et al.) through the use of "honesty"- a property meaning that the data used to grow a tree is not the same as the data used to generate fitted values on the tree. Honesty generates several desirable properties for the causal forest compared to other random forest methods, notably asymptotic consistency for the variance matrix.

The "generalized random forests" which are proposed in Athey, Tibshirani, et al. are "generalized" because they can be introduced to a variety of problems where the correct recursive partitioning method may vary. This method unifies a plethora of different research needs under one general framework which may be regarded as an adaptive nearest-neighbor algorithm (for my purposes of conditional mean estimation, the "ensemble learning" and "adaptive-NN" viewpoints are equivalent), thus alleviating traditional difficulties in understanding the properties of random forests[10]. The trees which are grown for my purposes are "causal trees" (Athey and Imbens 2016) and their ensemble a "causal forest" (Wager et al. 2018). Causal trees are adaptations of the standard CART algorithm which differ in their choice of splitting criteria and utilize a different objective function. Causal trees attempt to maximize the criterion of honesty, which is the expectation of mean-squared error of treatment effects, taken across the estimation sample, the test sample, and a training sample at every parent node. CART, by contrast, does not have a distinct training and estimation sample, just a training sample. From these estimations of treatment effects at every node of the causal trees in the forest, an individual treatment effect, $\hat{\tau}_i$ is obtained as the average of estimated treatment effects across all trees where the observation was "out of bag". That is to say, the individual estimates which are averaged out to obtain an ATE are the weighted averages, across all trees, of the values assigned to each leaf into which an observation falls when dropped "out of bag" down a tree. The use of the phrase "causal" might appear a bit salacious here. After all, I do not have access to a randomization mechanism, the treatment (whether or not the respondent has internet in their house) may seem dubious, and it's rather obvious that there are SUTVA violations present in the data because the internet is, by definition, a network where agents can affect one another. These are all genuine issues with the kind of proper estimation which would be needed to be totally certain of the effect which the internet has had on polarization. I offer two defenses of my method against these valid critiques. First, obtaining asymptotically valid estimates for treatment effects in this context requires the relatively relaxed assumption of conditional unconfoundedness, that is, $W_i \perp \{Y(0), Y(1)\}|X$ in the notation of the Rubin causal model (see, e.g., Rubin 2005), which I interpret to mean that, even if the estimation is noisy, the variances are asymptotically valid[11]. Whether this condition is truly upheld or not is unknowable and untestable, but I believe that my feature set is sufficiently rich to justify this assumption.

[10] There are many papers which have attempted to understand the properties of random forests in the past. Biau et al. 2016 provides coverage of some of these issues up to the paper's publishing, and Athey and Wager proves that the generalized random forest does not suffer from these same pitfalls

[11] Indeed, the relatively small sample size works against me here, providing larger than desired variance estimates and making the task of showing treatment effects more difficult

Second, I take great care in this work to talk about the effects that the internet generally had, but I am unable to pinpoint the particular channels through which the internet would have treatment effects on polarization. I cannot, for example, distinguish between a treatment effect which would be due to increased exposure to traditional media outlets and one due exclusively to fake news viewed on social media. I also cannot condition my results on the intensity of treatment because I do not have information on respondents' frequency of internet usage or media diet. I would expect, were I to have this information, that I would a positive correlation between my polarization measure and internet usage, and so I do not speak of my results in terms of intensity. Having established these caveats, this thesis's use of the internet as a treatment comes from the view that I can regard the introduction of the internet as a shock which is revelatory of the ways in which partisanship was already operating. I am testing the idea that this shock reveals patterns which are also present in data on stereotyping. These two streams of results, taken together, are suggestive of causal pathways, but not proof.

The results I report from ANES stem from six different causal forests, which differ in the choice of dependent variable and restrictions placed on the data set. The first set uses all respondents and its dependent variable is a combined measure of "thermometer" ratings for Democrats and Republicans, defined as $\frac{(A_i^D - A_i^R)+100}{2}$, where A_i^D and A_i^R denote the thermometer scores for Democrats and Republicans, respectively. The use of the combined measure is motivated by the fact that it can incorporate the views of self-declared independents. Although polarization would usually be thought to affect the "true" partisans solely, including independents has the upside of describing more general trends in polarization. One anticipated flaw with this methodology is that estimating the treatment effect for this outcome would be biased towards zero, because the treatment would move Republicans and Democrats in opposite directions from one another, and make treatment effects difficult to measure. It would also be unclear from just this measure alone to know if effects are driven by a greater negative affect for the opposite party (so-called "negative partisanship") or by increased affect for one's own party, and the inclusion of independents would only muddle such interpretation. However, were this forest to show a treatment effect, it would imply that the internet had a stronger "pull" factor towards the orthodoxy of one party. To better understand general polarization effects without this "tug of war", a second forest is implemented which uses an individualized version of the Boxell et al. measure as follows: $(A_i^D - A_i^R)\mathbb{1}(i \in D^L) + (A_i^R - A_i^D)\mathbb{1}(i \in R^L)$ where D^L and R^L indicates association with the Democrat and Republican party respectively, including independents who lean towards those parties. Thus, rather than having a yearly measure as in Boxell et al. 2017, I obtain individual-level observations of affective polarization. This measure corrects the problems with the first dependent variable by doing more to ensure that all treatment effects move in the same direction, hence I call it a "corrected affective polarization measure". From this measure, however, it is still not possible to know whether effects are due to negative partisanship or positive in-group affect. To obtain credible estimates for the level of "negative partisanship" which is occurring, the final four implementations look at affect towards out-parties: Democrats (excluding leaners) on Republicans, Republicans (excluding leaners) on Democrats, and independents (including leaners) on both separately. In this way, we can explore the role that the internet has played on negative partisanship for individuals who fall into specific partisan identities, with independents considered to better understand baseline trends in party-specific polarization based upon the internet. Table 1 presents a breakdown of the observations in each forest, subdivided according to a number of demographic variables which will be important for analysis later.

Table 1: Weight-Adjusted Relative Frequencies in Each Forest, by Demographic Group

Variable	Baseline Forest (n = 1716.77)	Corrected Affective Polarization Forest (n = 1625.52)	Democrats on Repbulicans (n = 535.82)	Republicans on Democrats (n = 515.13)	Independents Forests (n = 582.30)
Male	53.68%	53.22%	47.33%	55.00%	58.3%
Female	46.32%	46.78%	52.67%	45.00%	41.7%
By Year					
1996	9.7%	9.4%	13.2%	7.4%	8.9%
2000	15.6%	15.6%	16.8%	12.4%	15.9%
2004	28.4%	28.4%	25.0%	34.7%	27.7%
2008	21.4%	20.9%	21.2%	19.8%	23.5%
2012	24.9%	25.8%	23.8%	25.6%	23.9%
By Age Groups					
30 and Younger	21.3%	21.6%	22.6%	16.0%	24.7%
31 - 50	41.0%	40.3%	39.9%	41.6%	40.1%
51 - 70	29.1%	29.5%	28.3%	32.6%	27.7%
Older than 70	8.5%	8.5%	9.2%	9.8%	7.6%
Ethnicity					
White	76.3%	76.4%	60.1%	90.0%	79.8%
Black	10.2%	10.1%	23.2%	0.2%	7.2%
Asian or Pacific Islander	1.9%	2.0%	1.8%	1.6%	2.2%
American Indian or Alaska Native	0.8%	0.8%	0.8%	0.6%	1.0%
Hispanic	8.2%	8.3%	12.7%	6.0%	5.6%
Other or Multiple Races	2.6%	2.4%	1.4%	1.6%	4.3%
By Education					
Grade School	1.3%	1.3%	2.0%	1.3%	0.8%
High School	33.4%	33.0%	35.6%	29.6%	35.3%
Some College	31.5%	31.4%	30.1%	30.3%	32.6%
College or Advanced Degree	33.8%	34.2%	32.3%	38.8%	31.3%
Family Income Percentiles					
Bottom 16th	11.7%	11.2%	16.0%	6.3%	12.0%
17th to 33rd	14.7%	14.9%	17.0%	11.8%	14.9%
34th to 67th	37.3%	37.6%	37.3%	36.5%	39.1%
68th to 95th	28.9%	28.7%	25.1%	33.6%	28.9%
96th and Above	7.4%	7.6%	4.6%	11.7%	5.2%
Employment Status					
Employed	68.7%	68.6%	68.3%	69.3%	68.1%
Temporarily Laid Off	1.0%	1.0%	0.6%	0.6%	1.6%
Unemployed	3.0%	2.8%	2.4%	1.4%	4.6%
Retired	15.2%	15.4%	16.7%	18.5%	12.2%
Permanently Disabled	2.8%	2.7%	3.1%	2.3%	3.1%
Homemaker	6.7%	6.8%	4.7%	6.6%	7.9%
Student	2.7%	2.6%	4.1%	1.4%	2.6%
Marital Status					
Married	59.7%	59.5%	55.6%	71.4%	53.8%
Never Married	17.4%	17.4%	18.7%	10.9%	21.4%
Divorced	9.4%	9.2%	9.9%	6.7%	11.4%
Separated	2.7%	2.7%	4.0%	1.1%	2.4%
Widowed	6.6%	6.6%	6.3%	7.2%	6.7%
Partners; Not Married	4.3%	4.5%	5.5%	2.7%	4.3%

Weights are given by the variable VCF0009x in the ANES cumulative data file set

Although the variable sets that enter into each algorithm are exactly the same, the final algorithms' data sets will vary due to parsing that happens along the way. First, two separate regression forests are fit to the dependent variable and to the treatment, thus obtaining vectors of estimates $\hat{Y}_i$ and $\hat{e}_i$, the propensity score. With these two in place, a "raw" causal forest is run, which takes all of the information from the original data set[12]. This forest is used to understand the components of the data which are important for placing splits, contingent upon the dependent variable and the treatment variable. "Importance" in this context both signifies the conventional meaning of the word, as well as the meaning specific to random forests. In generalized random forests, importance is defined as a weighted sum of the number of times a variable generate a split, weighed inversely by the depth of the node. This measure thus rewards variables which are used early on, i.e., which provide effective splits for large amounts of data. The "real" causal forest[13] is an ensemble of 10,000 trees which is grown by restricting the data set to the variables which proved to be the most important[14], and by tuning parameters for the minimal node size, the imbalance penalty (how harshly the algorithm penalizes splits which result in a large imbalance between the number of observations in the two child node), the maximal allowable imbalance within a split, the honesty fraction, and the number of variables sampled at each node. Although there are more tunable parameters in the algorithm, a limited number had to be selected due to feasibility of parameter estimation given the small sample size. Thus, the out-of-the-box options are used for the sample fraction (i.e., the size of the initial data set for each tree) and, following the recommendation of Tibshirani et al. 2019, the option to prune leaves after fitting is turned off to improve performance with relatively small samples. In every case when tuning is performed, the algorithm solves for the parameters of interest by using cross-validation. All forests are grown with sample weights given by the ANES data, with standard error estimates clustered by year, and estimates reported are both heteroskedasticity consistent (with HC3 variance corrections used, following work in Semenova et al.) and cluster robust.

4 Analysis

Table 2 reports the average treatment effect estimates for the six causal forests as well as omnibus tests for the presence of heterogeneity. The omnibus test computes a best linear fit (in the sense of Chernozhukov et al.) for a regression of the residual terms of the forest with two regressors: the mean forest prediction provides the measure of "no heterogeneity" and the difference between the fitted value and the average fitted value (the so-called "differential forest prediction") provides evidence of heterogeneity. A coefficient for the "differential forest prediction" which is significantly different from zero is evidence of the general presence of heterogeneity in the data. Naturally, however, lack of significance is not evidence of a lack of heterogeneity, and Athey and Wager highlights that there may still be heterogeneity along particular variables of interest within the data. The table reveals a few baseline facts for the effect that the internet had on political polarization. Interestingly, the only significant values that come out of these estimates are ones from forests wherein the dependent variable is attitude towards Republicans. A

[12] In total, the original data frame is composed of 215 variables which is transformed into a matrix of 595 features once factor variables are turned into dummies

[13] When reporting the results of causal forests in this thesis, I am always referring to these "real" forests

[14] Following the previous papers on this method, the most important variables are defined as the variables whose importance measure was greater than the mean importance measure of the variables in the raw forest

negative value in that context signifies that sentiment towards Republicans worsened over this period. These results signify that, on average, non-Republicans felt colder towards Republicans due to the introduction of the internet[15]. These tests also fail to reject a null hypothesis of no heterogeneity (and no value really comes close to rejection) and most of the tests also fail to reject a null hypothesis of appropriate calibration for the mean forest prediction. Taken together, it seems that there is neither particularly strong evidence for a well-calibrated "mean" forest nor for a purely differential one in any of these models, save the ones where Republican affect is the dependent variable.

Table 2: Average Treatment Effect Estimates for Each Causal Forest

	Average Treatment Effect Estimates	Test Calibration for Heterogeneity	
	Estimate	Mean Forest Prediction	Differential Forest Prediction
Baseline Forest	0.3 (0.6)	0.10 (1.76)	0.44 (0.89)
Corrected Affective Polarization Forest	1.30 (1.61)	1.51 (1.19)	-0.03 (0.30)
Democrats on Republicans	**-5.69***** (2.02)	**1.17**** (0.60)	-0.14 (0.17)
Republicans on Democrats	-4.0 (4.08)	2.60 (2.33)	0.61 (0.60)
Independents on Democrats	-0.77 (6.22)	1.02 (0.92)	-0.13 (0.44)
Independents on Republicans	**-5.13*** (2.67)	**1.16**** (0.53)	0.16 (0.24)

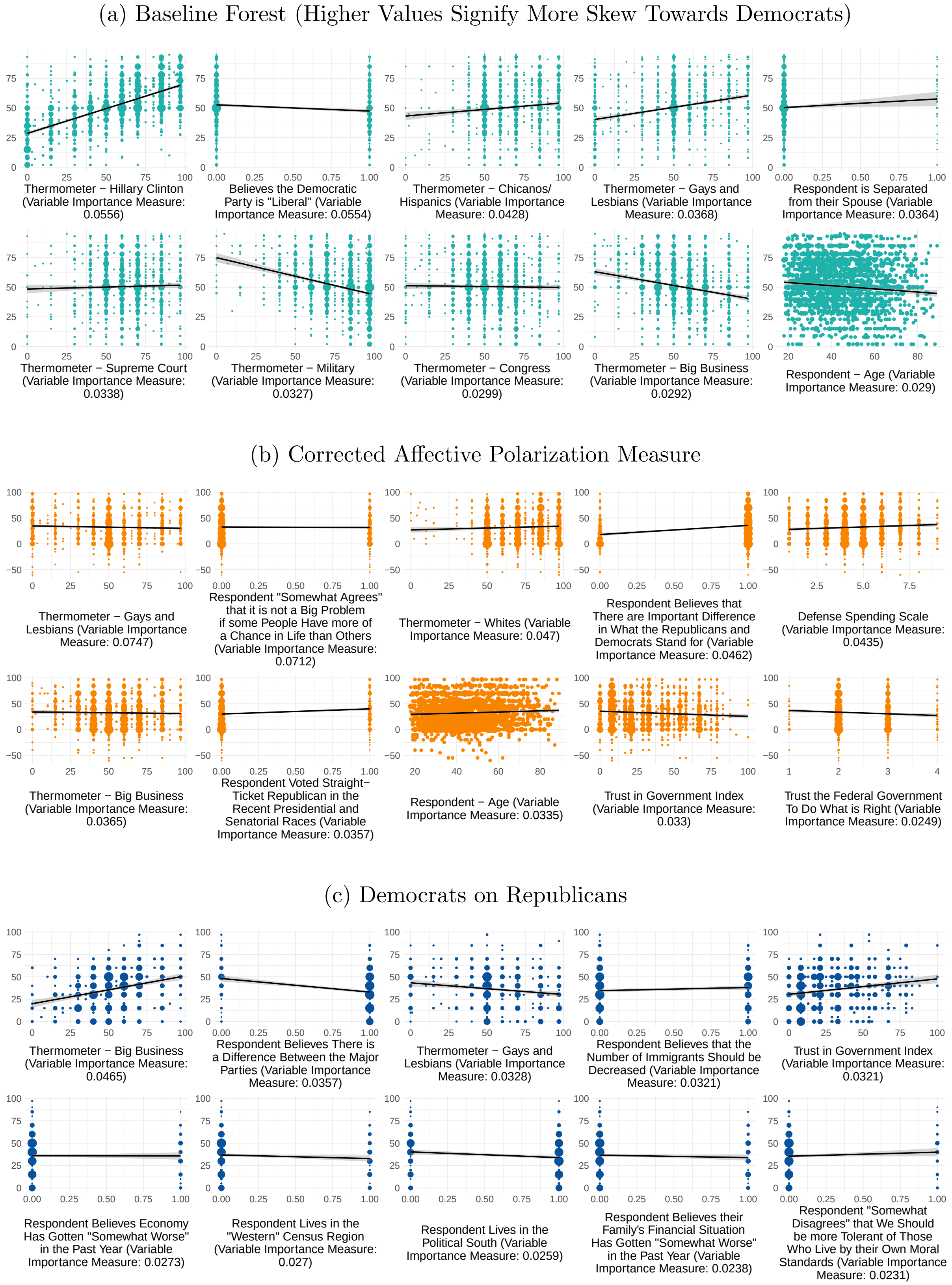
Values in Parenthesis Report Standard Errors
All Values Rounded to Two Decimal Places
* Significance at the 10 % level
** Significance at the 5 % level
*** Significance at the 1 % level

By way of a first approximation to understand how the forests are arriving at their fitted values and partitioning the data, I report the 10 most important measures for each forest in figure 2. It is impossible to say whether or not any treatment effect would be increasing or decreasing in any variable just by looking at this measure, but in an attempt to understand underlying relationships I report both the scatterplot of responses to this variable against answers to the dependent variable for each model, along with the OLS regression lines from a bivariate linear model between the two variables and the 95% confidence bands. Larger dots indicate higher (weighted) frequency of responses at the pair (x, y) in every case. When a variable in the grid is binary, a score of 1 corresponds to an affirmative response to the variable on the horizontal axis.

As mentioned in the previous section, a higher importance measure means that the

[15] A more stylized version of this information, though not a completely justified one given the data available, would be that the internet is a cold place towards Republicans, which broadly comports with contemporary GOP concerns around conservative censorship on social media platforms, for example.

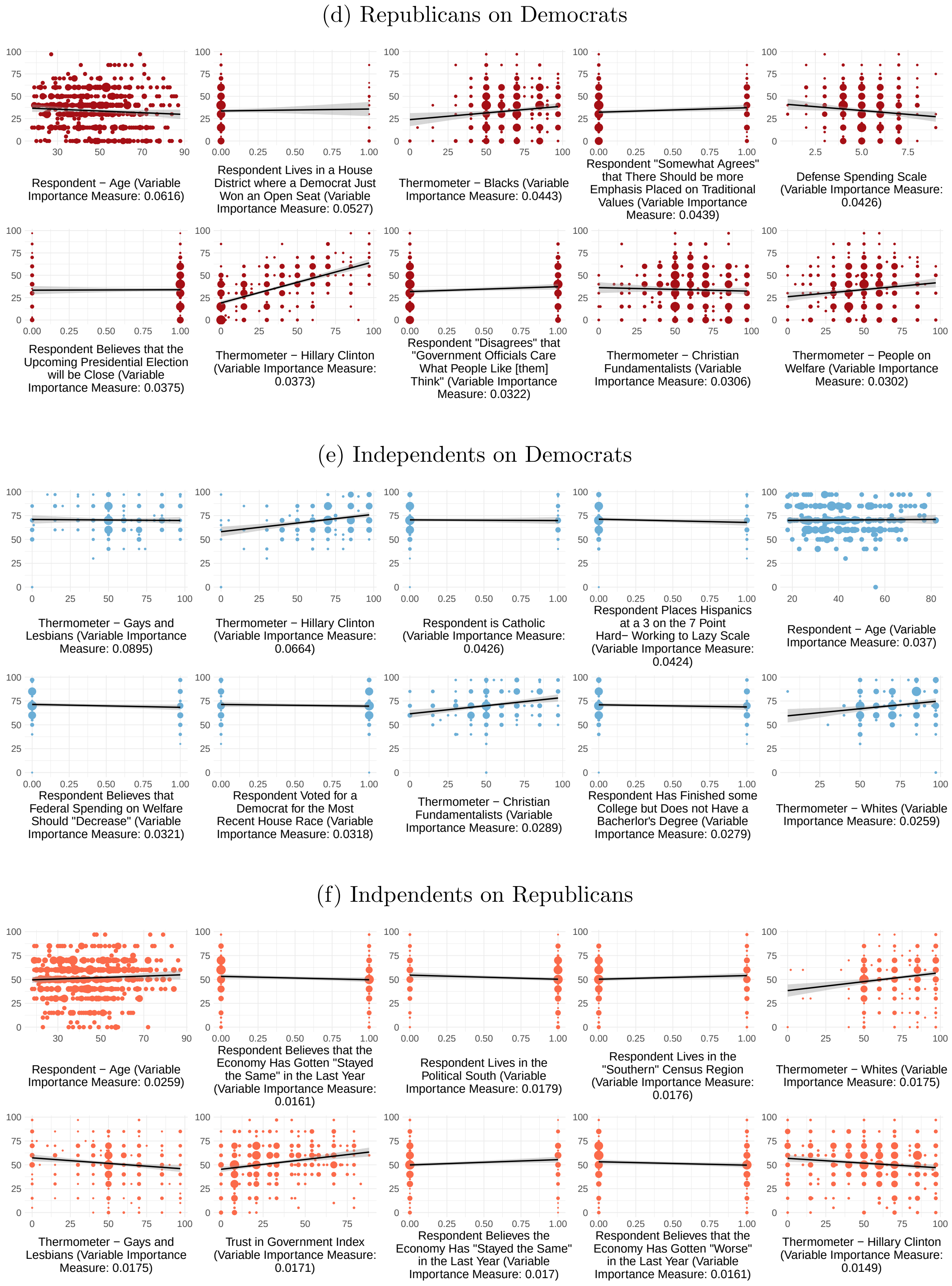
Figure 2: Bivariate Relationships Between Important Variables and the Dependent Variable, by Forest



(a) Baseline Forest (Higher Values Signify More Skew Towards Democrats)

(b) Corrected Affective Polarization Measure

(c) Democrats on Republicans



(d) Republicans on Democrats

(e) Independents on Democrats

(f) Indpendents on Republicans

algorithm utilized the variable more frequently when it was available as a choice compared to variables with lower importance measures. In these forests, the demographic variable which appears most frequently in the top 10 is the respondents' age. In essence, this demonstrates that every one of these algorithms utilizes age quite frequently, and often more than any other demographic measure. This gives me reason to believe that age is informative on general political disposition, and makes age a credible variable to test for heterogeneity. I would expect that age would be an informative direction for heterogeneity in each forest; while age does not appear in the top 10 for Democrats, it is just behind as the 11th most important variable. There are points of caution however. There may be more utility to the algorithm in splitting along age compared to any of the other demographic variables due to the particular structure of the data. Because age is a discrete variable along the range of 18-89 in the data, while most other demographic variables are coded as binary dummies, the algorithm has more choices along the age axis by which to split data, meaning that the age variable could be given outsized importance relative to binary choices. There is also reason to doubt this conclusion however; dummy variables were not wholesale excluded from being some of the most important which suggests that the algorithm is not discounting them entirely. Moreover, the fact that the causal forests are obtained after filtering out unimportant variables would imply that the algorithm has already identified the most important distinctions among binary dummy variables. For example, the causal forest using all respondents identifies that answering "worse" to the question "has the economy gotten better or worse in the last year" is relatively important for predictive accuracy. Within that data set, the algorithm also has the binary variable encoding a response of "no change", but not variables for "better" or "don't know" even though these variables were in the original "raw" forest. Thus, the algorithm has already found that the two surviving responses, which are neutral to negative on the question, were the most important for sorting responses in the context of the regression problem, and by removing the unimportant distinctions ex-ante, the final causal forest algorithm doesn't run the risk of "wasting splits" with an ultimately unimportant variable which just so happened to provide the best split at a given node, instead leaving them open to surviving variables from the first round. In the case of independents' opinions about Democrats, whether the respondent is Catholic or not is the third most important variable. My intuitive answer to explain this behavior would be the wedge issue of abortion, but without data on respondents' opinions on abortion, it is impossible to make this conclusion definitively [16]. It is also notable that the average sentiment towards Democrats among Independents is higher than their average sentiment towards Republicans. This is particularly surprising given that Democrats held an equal number of presidential terms compared to Republicans during this time, and perhaps speaks to the unpopularity of President George W. Bush towards the end of his second term[17].

[16] ANES does have recent questions on respondents' opinions of abortion, but these have been asked in inconsistent ways throughout the years, and had to be excluded from the overall data set due to said inconsistency in answers.

[17] A cursory glance at historical approval ratings for President Bush by Gallup Polls shows that his lowest approval rating ever attained was 25. By contrast, the lowest ratings attained by Presidents Clinton and Obama, the other two presidents during the time frame of this data set, were 42 and 41 respectively. Data available from https://news.gallup.com/poll/116500/presidential-approval-ratings-george-bush.aspx for Bush, https://news.gallup.com/poll/202742/obama-averages-job-approval-president.aspx for Obama, and https://news.gallup.com/poll/116584/presidential-approval-ratings-bill-clinton.aspx for Clinton.

Aside from age, a few other variables stand out. The fact that the respondents' attitude towards Hillary Clinton- who during this time was the First Lady, a Senator from New York, and the Secretary of State in the Obama Administration, but not the Democratic nominee for the presidency- is quite striking. Clinton held positions of relative national prominence during the time, but these positions were paltry compared to the national attention gained by her presidential run four years after my data ends. On face, its importance means that responses to this variable sort the various respondents quite well, which makes intuitive, if uncorroborated sense because (a) highly-informed partisans tend to be more polarized[18], and (b) polarized attitudes towards Hillary Clinton, who was not a major candidate during the course of this study, would be revelatory for partisanship in other ways, particularly because the association between high affect for Hillary Clinton and Democratic-leaning views is so strong. For the causal forests looking at Republican attitudes towards Democrats, one of the more interesting important variables is the opinion of people on welfare. This demonstrates early evidence in favor of the model of party cognition proposed by Ahler et al. 2018, where respondents think about party-stereotypical groups when asked about the parties.

Along similar lines, racial attitudes are also clearly important in creating useful splits. A thermometer rating for an ethnic group shows up in the top 10 for every forest except the one for Democrats on Republicans. Generally speaking, these measures correspond with the direction that we would expect given party-prototypical stereotypes, with the exception of the corrected affective polarization forest, where the measure doesn't permit such conclusions. In fact, it would seem that a linear trend isn't appropriate in that case, and that an upward-facing parabola would perhaps be a better fit in order to capture the extreme associations on either pole of the thermometer rating. The final group of particular interest here would be the gay and lesbian community[19], for which the thermometer variable appears in the top 10 of every forest except the one for Republicans on Democrats. That the variable is important is perhaps not too surprising; the US underwent significant changes in the average sentiment towards this community over the period of study. However, the fact that this variable does not show up for Republicans is quite surprising, as this community is one identified in Ahler et al. 2018 as being prototypical for Democrats. Further exploration is needed to identify if an anticipated intra-party heterogeneity along this variable is actually present in the data.

4.1 Heterogeneity by Single-Variable Splits

Although variable importance gives us a sense of a variable's split utilization, weighed by the depth of the node at which it was used, I push that measure forward by also examining the values which were used in splitting to hone into the role of heterogeneity in the data set. I use the information provided by the split data to divide individuals into one of two categories, and estimate the differences in ATEs between these two newly generated subsets. In the cases when a binary variable is used, this becomes an estimation for the difference in ATE between an affirmative response (coded one) and a negative response (coded zero). From the outset, testing on factor variables introduces quite large standard

[18] Subsequent portions of analysis demonstrate that polarized people sometimes eschew their political knowledge when assessing upcoming elections

[19] The use of the phrase "gay and lesbian" community or "lesbian and gay" community throughout this thesis is in accordance with the wording of variable VCF0232 in the ANES cumulative data file. It does not serve as shorthand for the broader LGBTQ+ movement and I keep the wording precise so as not to overstate the influence of the variable given its construction

errors on the average treatment effects because a response coded as zero could mean that a respondent is away from the initial value in either direction. For this reason, it is important to consider the levels of the alternative responses for the same variable when possible[20]. When the variable is an integer, I test at two different thresholds, the mean and modal values used to generate splits across the causal forest. Of these two, the modal value is perhaps the more methodologically strong, as the variable importance measure is defined by frequency and depth of splits, from which I interpret modal splits as being particularly "important" values in driving more general variable importance. There may also be an objection at this point to the use of what are, in some sense, seemingly arbitrary threshold values for these variables, akin to the problem of false thresholds in a regression discontinuity context. After all, just because I find a variable to be significant at one value, that does not necessarily imply that the variable wouldn't be significant elsewhere. The logic also works the other way: even if I find insignificance in one place, that would not be a reason to discount the variable (which may be critical to the data generating process) wholesale. To this, I would point out that the algorithm is attempting to maximize heterogeneity when it creates a split, though it is certainly the case that significance in one place does not imply robust significance, nor does it imply insignificance elsewhere. To engage with this concern throughout the section that follows, I treat the variables which are significant at some threshold as evidence of underlying cleavages within the population that also occur at that threshold, rather than treat the threshold values as "magic numbers" which the algorithm happened to find[21]. It is still nonetheless the case that these results are just looking at one variable at a time. However, I would expect that subgroup-based differences in polarization, if they exist, would do so in a region defined by multiple variables. Having said all of this, the basic logic of this exercise is to check whether or not the split values which the algorithm are selecting in particular trees are informative about the existence of heterogeneity in the entire forest. In essence, then, what follows is a discussion of the divisions which the algorithm made frequently, and which are highly informative about the structure of division within society.

As expected, few variables in the data sets demonstrate significant treatment effects between high and low regions. There are, however, some exceptions that are worth discussing for each forest. In the baseline forest, we see encouraging signs that stereotyping is a primary for political polarization. Variable VCF92714 asks respondents to place Americans who are black on a scale from one-to-seven, where seven signifies that the respondent believes most Americans who are black are lazy, one signifies that the respondent believes most are hard-working, and four signifies that the respondent believes most are not close to one end or the other. Because this variable is a factor, the levels of possible responses are each given their own variable, though not every level enters into the analysis because the forest is grown only with levels which were important in the raw forest. Here, the

[20] To use an example which will not show up in any of these sets, if a respondent indicated that their mother was "somewhat interested" in politics, then the binary variable created around that factor would include a value of zero which captures both individuals who said that their mother "did not pay much attention" and those who said that their mother was "very much interested" in politics. Thus, assuming that answering this question has serious bearing on the variable of interest, the standard error estimates for not answering "somewhat interested" depend on the relative proportions of both answers which are not "somewhat interested". This imbalance must be considered when interpreting significant results from head-to-head comparisons using binary variables

[21] For example, if I were to find evidence of heterogeneity in the population along the variable covering affect towards white people at the value 50, then I would be led to conclude that the bifurcation which occurs at or around 50 is important, not that answering above or below 50 is the critical threshold for how relatively polarized one is.

levels corresponding to a “lean” towards hardworking and lazy (i.e., values three and five on the scale respectively) both enter into the algorithm, as does the neutral response value of four. Answering this question with neutrality, or with a slight lean towards laziness are both significant variables, while answering with a lean towards hardworking is not. Moreover, the signs move in the opposite direction of one another, implying that there is a sharp break point between the type of individuals whose answer is in the middle and those whose answer leans towards laziness. One immediate interpretation of this result would be that respondents generally choose to place themselves somewhere around the middle of this scale (a social desirability effect would imply that very few people choose response seven) around which to coalesce, and the side to which they lean is informative of their partisan identity. If this interpretation is correct, then there is a novel difference between the party’s averages on this scale which are evidence of stereotypical thinking but not definitive proof. This evidence is bolstered by the fact that the thermometer rating for Americans who are black is also a significant variable in this iteration of the forest, and also exhibits a sharp drop off in the number of responses lower than 50. Around the mean split value of 61.7, individuals who fall to the right exhibit a four-point increase in their net thermometer score. Finally of note, being a homemaker is a significant variable in three different factor variables which capture employment responses, all of which have a positive estimate[22]. The most parsimonious explanation, to be revisited later in this section, is that the homemaker variable is actually capturing a treatment effect for women, who make up 94.7 percent of homemakers in this data set. I believe this to be the case in particular because the gender variable was removed from this forest after the first pass, and so the homemaker variable becomes a useful measure by which to split a subset of people who are overwhelmingly women from the rest of the population. In short, with a noisy model that cannot properly distinguish between love of one’s own party and hate for the other party, the variables which stand out from the rest are those which capture racial attitudes and respondents’ gender.

In the corrected affective polarization forest, there are only three variables which demonstrate significance. The first is a measure of the thermometer towards the military, which has an estimated 10 point difference between those whose rating was above 85 and those whose rating was below 85 (the modal number used for splits in the causal forest). Upon closer inspection, while there are certainly individuals from both political persuasions (i.e., Democrat and Republican leaning) who feel very warmly towards the military, there is a large difference in the numerosity of each group above 85. Of those who answered above 85, 61 percent are Republican-leaning. Due to the complicated nature of interactions between multiple variables as described before, it is of course quite possible that the significance of this variable is an informative signal on social cleavages in some other dimension which has little to do with politics. For example, it could be the case that the types of respondents who hold a strongly positive affect for the military do so in part because they work with the military in some capacity. What seems most vexing about the

[22] It might seem quite strange that there are three different employment responses present in the data which get entered into the algorithm. ANES contains variables which capture employment according to a number of different possible categories, meaning that each has a differing level of granularity. Having each level of granularity in the algorithm is somewhat desirable because it allows for employment variables to enter into split consideration more frequently and, in cases where multiple employment variables enter into split consideration simultaneously, allows the algorithm to pick the level of aggregation in employment which is the best to split the node. The fact that all of the estimates for this homemaker variable tend to be around the same value is encouraging because this population should be fairly consistent between all of those measures.

significance of this variable, and the reason why I choose to interpret significance here as evidence of other differences, is the difficulty in understanding why being especially warm towards the military should mean that an individual becomes 10 points more polarized when introduced to the internet. There are potential avenues of interpretation dealing with in-group loyalty (for example, the kind of people who rate the military highly are the kind who believe in having cohesive teams like the military, and so exhibit a higher level of in-group preferences). However, the fact that these individuals make up 33 percent of the sample population makes me skeptical of the idea that these respondents all display a kind of in-group loyalty such that this would be a systematic effect present for all.

The next variable, that the respondent believes the election will be close, also has a puzzling element to it, namely that the difference in treatment between the two groups is negative. This mean that individuals who believe elections are likely to be close are, on average, less polarized than those who believe that elections will be easily won by one candidate or another. This result contravenes predictions of Downsian electoral competition, because those models would posit that individuals who are polarized (and are thus more motivated to engage in national politics despite the fact that their influence on national politics is low) would be doing so in part because they believe that upcoming elections will be close. While I cannot prove this definitively, it would seem that the significance of this variable is an artifact of a polarized media diet. Using correct identification of the party with the House majority as a proxy for baseline political information, the average score for the affective gap between the parties is highest for the group of voters who believe that the election will not be close and who are baseline "informed", and among the group of voters who say the election will not be close, over half (51.4 percent) are baseline informed. I reason that this is a group of individuals who, despite their relatively decent knowledge of American politics, still engage in highly motivated reasoning. The fact that this information seems to only make the problem worse also aligns with the findings of Ahler et al. 2018 where, for every party-group dyad except rich Republicans, the direction of misperceptions was an increasing function of interest in political news.

Examining the final four forests of partisans on opposing parties, most of the variables which are revealed to be significant are concerned with matters of moral philosophy or of government policy compared to status quo. For example, there are differences in the sentiment towards Republicans between Democrats who somewhat agree that it is a big problem if not everyone has the same chance in life and those who don't (by an estimated 14.4 points). Because we're just examining the rating towards Republicans, a positive value here means that the individuals who somewhat agree are warmer towards Republicans than those who do not. Although the most common response to this question beyond somewhat agree is "strongly agree", implying that those individuals are, on average, more hostile towards Republicans than those who only somewhat agree, the strongest difference between this group and the rest is the warmth that respondents who did not agree show towards the Republican party. This result, too, makes sense when considering that those individuals are placed in contrast between their identity as Democrats and their philosophy which runs contrary to the most common disposition of Democrats (indeed, those who agree in some capacity with this statement make up nearly 64 percent of the party). When such individuals face conflicts between multiple aspects of their identity, for example between their party and a political philosophy which puts them at odds with their party, they are less likely to maintain the "average" position of their party as that party moves to an extreme. These individuals, in short, do not have "mega-identities" Mason 2013 - a strong correlation between the political and social identities that an individual has which “stack” upon one another - which reinforces their political biases according to the coincidence of their political and personal identities or philosophies.

This same behavior is present in the question concerning whether we should be more tolerant of people who live according to their own moral standards. The difference in treatment effect for individuals who somewhat disagree with this sentiment compared to the rest is 19.9 points, indicating cross-cutting pressure between a viewpoint which is more commonly attributed to Republicans and the respondent's identity as a Democrat (Around 15 percent of Republicans and 8 percent of Democrats somewhat disagree with this statement, and most Democrats are concentrated towards somewhat agreeing, while Republicans demonstrate bimodality on this question with peaks at somewhat agreeing and somewhat disagreeing). In this data set, we also see that the measure of correctly identifying the party with a House majority (my measure of baseline information above) is significant and negative, meaning those who have this baseline level of information feel colder towards Republicans than those who do not[23]. When Democrats indicated that they were “very much interested” in the upcoming election, the treatment effect difference compared to the rest is estimated at -9.77 points, significant at the 5% level. Again, this finding is in line with what we might expect: those who are paying close attention to what is happening are the kinds of people who, upon gaining internet access with the variety of media sources available, sort themselves into a media ecosystem which reinforces their views and makes them feel comparatively colder towards opposing viewpoints.

For Republicans, the only significant variable from this initial analysis concerns views on child care. Respondents who believe that federal spending on child care should remain the same have a near 20 point difference in treatment compared to those who don't (significant at the 5 % level). This finding is, however, somewhat dubious; the majority of remaining respondents believe spending should be increased and have higher ratings towards Democrats by about 11 points on average, while respondents who believed such spending ought to be decreased rated Democrats less favorably by about the same amount, and those who were unsure (a weight-adjusted 8.69 people out of 515.13) have lower ratings than those who want less spending by 17 points on average. These extreme outliers are likely driving the result as the algorithm estimates that being a member of the non-status-quo group corresponds to a treatment effect of -11.34 points, despite the relatively even balance between average scores on either side and the higher numerosity of those who wanted increased spending.

For independents, the variables which stand out concern demographic issues, similar to the noisy baseline forest. For the rating of independents on Democrats, the only significant variable is the thermometer rating on gays and lesbians, which is significant when the data is separated between the modal split value of 15. Those on the right of this value have a treatment effect difference of 20.4 points (significant at the 5% level) compared to those on the left. The fact that attitudes towards gays and lesbians has such a strong treatment effect may be indicative of the kinds of stereotypes that these independents believe about the Democratic party. There are, broadly, two pieces of evidence which corroborate this line of reasoning. The first comes by way of Ahler et al. 2018, which finds that survey respondents egregiously overestimate the proportion of Democratic party members who are lesbian, gay, or bisexual[24], believing that the share of lesbian, gay, or bisexual

[23] As a sanity check, an inspection of the average rating towards Democrats between these two groups also demonstrates that the uninformed feel warmer towards both groups, with a difference between parties which is higher by a margin of two points when compared to the informed

[24] While it is not the case that a question about the lesbian and gay community directly maps on to a

Democrats to be five times higher than it is in reality (31.7% compared to 6.3%). Hence, on average for independents, negative feelings about the lesbian and gay community can be expected to map more strongly onto feelings about the Democratic party than would be expected from the actual proportions of LGBTQ members in the Democratic party. This is particularly true given that 15 is a remarkably low value for this group, whose average rating is 54.38. From this, I infer that heterogeneity along this variable is capturing a rather insular group of individuals with highly negative views towards the gay and lesbian community. There is yet more to say here. By dividing the group of independents according to their partisan leanings, I find that the average rating for the lesbian and gay community is 55.93 among independents who lean Democratic, 44.12 points among "true" independents, and 44.83 points among those who "lean" Republican. Moreover, the average ratings for Democrats in these three groups are (respectively) 64.36, 53.87, and 41.24. I would argue that the reason for this decreasing warmth could also lie in the increasing social distance between the groups. That is to say, because the lesbian and gay community is representative of the Democratic party (in the sense of stereotypes a la Gennaioli et al. 2019 or party prototypes a la Ahler et al. 2018), negative views towards this community map more strongly onto the Democratic party the further away one is from affiliating with the Democratic party. We see preliminary evidence of stereotyping which we can verify with the Ahler et al. 2018 set later on in the analysis.

When asked to rate Republicans, there are four variables which show heterogeneity when split: a binary variable indicating if the respondent places Americans who are white in the middle of the hard-working to lazy scale, whether there's a Senate race in the state, whether or not the respondent is opposed to affirmative action in hiring and promotion processes, and (the follow-up to the previous question) whether or not the respondent is weakly opposed to affirmative action in hiring and promotion. Here, too, then we see that variables which ask about relationships to ethnic groups enter the analysis as a source of heterogeneity as they did for the baseline forest. However, unlike the linear relationship that was observed between the thermometer rating for Americans who are black and the dependent variable, here the average rating for Republicans is nonlinear. The average feeling towards Republicans reaches its lowest point at a hardworking-lazy rating of 5 (the response corresponding to a belief that some small majority of whites are lazy), its second lowest point at 4, and showing large departures for every other value on the scale[25]. On the affirmative action questions, neither of the measures reveal particularly stark or unexpected differences between the two partisan persuasions. However, because the proportion of survey respondents who fall into generally opposing affirmative action is quite large (81.1 percent of the entire sample and 82.5 percent of declared independents), there are likely some important differences between those who are in favor and opposed. On this point, while explanations on the differences between respondents may not be found in political affiliation, there is slightly stronger evidence that the differences are expressed along racial lines. Among nonwhites, 34.9 percent answer that they are not opposed to affirmative action (compared to 13 percent of whites), and nonwhites make

question about lesbians, gays, and bisexuals, or indeed the broader LGBTQ+ movement, I would expect that the overlap will be similar enough for this point to still be true, even if by a lower magnitude

[25] There are caveats to extrapolating information from this trend, however. For one, the weighted sample sizes become minuscule as one moves towards seven on the scale, such that the averages presented are really the combination of three to four survey respondents. Somewhat anecdotally, at a value of 6 on this scale, the single Republican-leaning respondent increases the average by placing Republicans at a score of 70 on the thermometer scale, a rather unexpected value given the predictions of representativeness, but such a case is obviously too insular to make large conclusions about the theory more generally

up 45 percent of the total responses in favor of affirmative action[26].

4.2 Heterogeneity Based On Demographic Variables

In this section, I utilize the results from the single variable splits in the corrected affective polarization forest, as well as differentiation by estimated treatment effects and general party differences to project heterogeneity onto variables of interest. The process of projection takes the treatment effect within a subset specified by the researcher, and approximates a doubly robust[27] linear model where the intercept is the baseline treatment and the independent variables are the researcher-specified variables of projection. Although the coefficients reported in the following tables There are a few different motivations for this methodology: this so-called "heuristic" approach[28] of identifying treatment effect heterogeneity based on estimated treatment effects comes by way of Athey and Wager. In essence, this method accounts for the fact that there are different regions of treatment effects present in the data, and the process which governs whether or not heterogeneity exists may be different in each region. The upshot of this method is that I need not assume that heterogeneity is only a process which effects the most polarized (or "anti-polarized") individuals. I would further argue that in a relatively small data set such as this, examining one subset of the data at a time may also inoculate portions of the data from the noise of other regions, allowing me to focus on the heart of the problem: identifying treatment effect heterogeneity along demographic variables. The costs of implementing this method are twofold: a loss in point estimate precision and a weak claim on external validity. The loss in precision is particularly acute when considering that the data set being used here is made up of 1586 observations, which become quite small as one looks at obscure slices of the population (e.g. all temporarily laid-off Democrats who have a relatively low treatment effect estimation). Prima facie, this implies that there will be large standard errors associated with estimates of the coefficients of projection. In truth, the following tables do demonstrate that the projection estimates can be quite unstable and, in particular, that the estimates associated with the intercept terms, which are often large, switch signs frequently[29]. While I will perform the analysis based around the estimated treatment as in Athey and Wager, I first proceed with analysis that sorts respondents based on whether or not they thought the upcoming election would be close. Doing this utilizes the information which was learned in the last section, i.e., that het-

[26] Some of these comparisons become relatively more or less favorable if one further subdivides nonwhites as in ANES survey questions VCF0105a and VCF0105b. In those cases, over half (56.8 percent) of respondents who are black are in favor of affirmative action, while only about a quarter (27.8 percent) of respondents of Asian or Pacific Islander descent are in favor. The white-nonwhite distinction is useful for comparing larger groups within the survey than is allowed by looking at each racial group atomistically, but subsequent efforts at estimating racial differences will not aggregate to the level of white and nonwhite both because it should be unnecessary when looking at the entire sample rather than independents, and out of the belief that such a comparison is not a useful grounds upon which to estimate differences in stereotyping

[27] Following Semenova et al., these estimates are doubly robust, meaning that they hold even if either the propensity model or the regression model is incorrectly specified

[28] This is the way that the original paper using this method frames it. Calling this method "heuristic" does not, however, refer to heuristics in the sense of Kahneman 2013

[29] Following the documentation for the grf package used to implement this method (Tibshirani et al. 2019), the intercepts could hypothetically take any value, but should be the ATE for the given subset when the features used for projection are mean 0. Because the demographic variables which will be used quite often in this section are not, it is difficult to say with certainty whether the resulting estimates are evidence of a poorly calibrated model or not

erogeneity in some form exists in the regions delineated by responses to this question, while also preserving the spirit of analysis coming from Athey and Wager, and adapting that method to the problem-specific regions of heterogeneity which have been previously identified.

Table 3 shows the results of four different iterations of the "best linear projection" function (Athey, Tibshirani, et al.) which differ in the choice of variables utilized for projections and sample subsets examined. The first column shows the differences among political parties in the entire subsample of individuals who indicated that they believed the election would be close, the second presents results on the same subsample while projecting the CATE to a suite of demographic variables, while the third and fourth project along the demographic variables with attention restricted to Democrats and Republicans (including "leaners"), respectively. I choose to look at parties, then demographics in order to answer two distinct, but related questions. The first is whether heterogeneity that is observed between the group who answered "close" and those who didn't can be sensibly framed as a distinction between the parties. In this case, it is clear that it cannot; the standard errors are quite large and the t-statistics associated with each estimate are tiny in absolute value. The second portion then examines whether those previously observed differences are occurring along demographic lines within each party. I conceive of this method as isolating demographic-driven changes which are occurring within each party and thus better identifying if a granular subset of one party demonstrates heterogeneity from the rest of the party, perhaps at the risk of implicitly assuming that social strata have more in common with fellow party members than fellow strata members (e.g. that rich Republicans are more like poor Republicans than they are like rich Democrats). As a sanity check to make sure that this potential assumption did not warp outcomes too much, I also performed the best linear projection onto the suite of demographic variables for the entire sample with party affiliation included. I found that the inclusion of party identifiers caused membership in the 96th-100th percentile of income to drop out of significance, but did not change anything else. Since I'm already reporting the projections for the parties independently as a way to understand the difference between partisans, I choose to keep it out of the specification with demographic variables.

Table 3 demonstrates that we do not observe strong heterogeneity along demographics in the region defined by believing the election will be close. Although there are certain standouts where heterogeneity is observed. Compared to the baseline lowest 16th percentile, the wealthiest individuals who believe the election will be close show a CATE estimate which, ceteris paribus, is 10.45 points lower. In addition, the homemaker variable which was a source of heterogeneity in the "noisy" forest also shows up as a source of heterogeneity among Democrats. This could imply that the effect which was observed there was concentrated among Democrats, but such a conclusion is not justified merely by the significance of this variable.

Given the inclusion of both a female indicator and the homemaker variable in this setting, my previous conjecture about the correspondence between the two causing significance would appear to be unjustified. Notice also that the two variables jointly demonstrate that the algorithm (nonsignificantly) estimates being a woman to be depolarizing compared to being a man, but (significantly) estimates that being a homemaker is a polarizing occupation. It could be the case that the high association (at least, one way association) between the two leads to poor estimation akin to multicollinearity in the traditional linear regression setting, but such a theory would need to be tested further. In addition to the homemaker variable, we see that the 17th to 33rd percentile indicator is

Table 3: Projections Among Respondents Who Believe the Election Will Be Close

Term	By Party Affiliation (n = 1285)	Among all Respondents (n = 1285)	Among Democrats (n = 709)	Among Republicans (n = 576)
Intercept	-0.65 (6.01)	17.47 (35.44)	-13.39 (35.33)	24.37 (19.05)
Democrats	2.78 (6.76)			
Independents	-1.13 (4.62)			
No Preference	18.98 (22.03)			
Other	-9.85 (31.57)			
Age		0.05 (0.15)	-0.29 (0.25)	0.39 (0.33)
Female		-1.70 (6.20)	-5.24 (6.62)	1.29 (7.80)
Ethnicity - Compared to Whites				
Black		1.25 (6.95)	-0.27 (9.17)	-8.58 (11.92)
Hispanic		6.31 (4.08)	-0.87 (5.68)	13.58 (13.25)
Other or Multiple Races		-9.49 (8.02)	-8.47 (14.31)	**-10.57**** (5.18)
Education - Compared to Grade School or Less				
High School		-16.49 (26.24)	28.40 (23.82)	-38.22 (24.33)
Some College		-10.24 (28.74)	36.95 (23.35)	-33.79 (21.44)
College or Advanced Degree		-12.58 (28.42)	36.51 (27.39)	-38.73 (24.09)
Income - Compared to the Lowest 16%				
17th to 33rd Percentile		-4.76 (8.74)	**-8.30*** (4.33)	-2.14 (17.63)
34th to 67th Percentile		-7.10 (6.65)	-3.85 (3.54)	-9.79 (16.38)
68th to 95th Percentile		-5.36 (8.98)	-5.59 (14.39)	-2.72 (17.57)
96th to 100th Percentile		**-10.45**** (4.94)	-13.76 (9.99)	-6.21 (11.63)
Employment Status - Compared to Currently Working				
Temporarily Laid Off		11.87 (14.85)	8.63 (13.06)	15.20 (18.93)
Unemployed		3.41 (15.84)	-4.66 (9.03)	14.34 (30.65)
Retired		5.24 (4.56)	14.37 (11.08)	-4.60 (12.56)
Permanently Disabled		2.55 (22.72)	-2.27 (20.41)	2.92 (21.98)
Homemaker		1.55 (5.06)	**19.38*** (10.26)	-10.45 (10.62)
Student		0.17 (7.05)	-3.92 (7.00)	0.72 (16.32)
Marital Status - Compared to Married				
Never Married		5.13 (5.49)	4.96 (9.03)	5.41 (13.42)
Divorced		-8.78 (8.06)	5.90 (8.70)	-26.68 (25.45)
Separated		-8.91 (14.21)	-15.95 (11.17)	15.08 (21.61)
Widowed		**-18.92*** (10.15)	-9.33 (18.61)	**-30.28**** (14.90)
Partners, Not Married		-1.02 (22.50)	-2.22 (26.00)	2.34 (8.48)

Estimates for parties use Republicans as the reference group
All values rounded to two decimals
Standard Errors shown in parentheses
* Significance at the 10% level
** Significance at the 5% level
*** Significance at the 1% level

also significant at the 10% level, and negative, implying that this group is less polarized than their poorer counterparts. Given the span of data in this set, such a relationship is perhaps unsurprising. Although recent elections (most notably the US election of 2016) have upended traditional alliances between the left and the poor (see also Gennaioli et al. 2019 for an example from France), the time frame covered here is one in which that alliance was strong, i.e. economic status was a highly informative dividing line between Republicans and Democrats. Hence, I would expect that the poorest individuals, those who are the most central in the party's message and have many people like themselves in the party, feel a high degree of in-group favoritism[30].

For the Republicans, the two variables along which I detect heterogeneity are the binary coding for being listed as "other" or "multiple" races and the binary coding for being widowed. Because this measure of ethnicity is not particularly nuanced (for example, it is less granular than other variables in the data set which could not be used because I could not guarantee that projection was possible for each desired iteration with a set of common variables across all of them), we cannot separate out the differences between mixed race individuals and Asian Americans. Regardless, the significance of the variable tells us that these individuals had a much lower level of affective polarization induced by the internet when compared to their white Republican compatriots. As in the previous section, my interpretation of a result like this is by virtue of cross-cutting identities: these individuals demonstrate less sorting behavior than their counterparts because they have more identities in disagreement. As a direct result, they are less polarized when a shock to polarization is induced as by the internet[31]. For the widowed in table 3, columns two and four, I am again tempted to interpret results similarly to the homemaker variable for Democrats, i.e., as moderating deeper results having to do with another, more fundamental variable. As in that case, such a conclusion is unjustified and is in fact less justified than the previous. Among the widowed respondents, 60% are 65 or above (compared to the 94.7% of homemakers who were women), so although there is a high degree of association between the two, there is clearly something particular about being widowed which implies that its significance here is not an artifact of close association with age.

Next, I examine the individuals who believed that the election would not be close. In the previous section, I theorized that this was the group of respondents who were demonstrating highly polarized behavior due to a poor media diet. Using the measures of media exposure provided by ANES, I was unable to confirm whether or not this is the case. Of the questions dealing with media exposure that might be useful, three of them are unavailable past 2008, and so were not usable here. The one that can be used, which asks respondents about whether or not they watched TV programs concerning the election, has a positive point estimate (not pictured) of 13.10 points, but is not significant. So at the very least, I am not able to delineate between the types of exposure that these individuals have and, thus, not able to project treatment effect differences onto a variable capturing the quality of media exposure, though the prima facie analysis that I have carried out suggests that there is not treatment effect heterogeneity between those who watch any television programs about the campaign and those who don't. Further, as shown in

[30] In a way, this is akin to the argument I made in the previous section about the LGBTQ community, where lower social distance was a plausible explanation for the difference in out-group affect. If anything, such an explanation would be more resonant here.

[31] Here, it is important to note that the intercept is estimated at around 24 points. Understanding that this variable is not significant but taking it as given for the moment, this model would predict that, ceteris paribus, mixed race or "other" respondents are still polarized, just by less (much less in fact) than their white counterparts

Table 4: Projections Among Respondents Who Believe the Election Will Not Be Close

Term	By Party Affiliation (n = 297)	Among all Respondents (n = 297)	Among Democrats (n = 175)	Among Republicans (n = 122)
Intercept	**20.92***** (7.94)	-29.11 (44.26)	-8.46 (37.95)	71.83 (94.96)
Democrats	-20.11 (18.49)			
Independents	-12.12 (15.39)			
No Preference	**-27.99***** (9.19)			
Other	-19.36 (26.31)			
Age		0.39 (0.87)	0.32 (1.41)	-0.18 (1.48)
Female		-7.43 (24.94)	-5.44 (20.99)	2.78 (66.74)
Ethnicity - Compared to Whites				
Black		8.36 (26.86)	19.54 (24.86)	-38.46 (77.07)
Hispanic		24.29 (27.32)	17.44 (21.96)	54.15 (62.16)
Other or Multiple Races		36.31 (23.12)	16.48 (42.83)	47.56 (58.38)
Education - Compared to Grade School or Less				
High School		23.08 (41.90)		
Some College		24.58 (47.52)		
College or Advanced Degree		28.51 (61.55)		
Income - Compared to the Lowest 16%				
17th to 33rd Percentile		-12.99 (21.54)	-10.47 (14.67)	-30.15 (50.65)
34th to 67th Percentile		-7.37 (8.95)	6.77 (26.72)	-36.44 (32.29)
68th to 95th Percentile		-14.19 (27.42)	-4.82 (41.32)	-46.77 (44.77)
96th to 100th Percentile		19.03 (31.99)	3.31 (32.88)	24.93 (122.74)
Employment Status - Compared to Currently Working				
Temporarily Laid Off		-17.08 (115.05)	2.53 (185.54)	–
Unemployed		-33.61 (40.46)	-57.50 (104.88)	-12.68 (109.44)
Retired		2.05 (25.78)	6.36 (25.86)	14.71 (92.46)
Permanently Disabled		15.09 (21.96)	21.53 (40.93)	29.02 (50.30)
Homemaker		15.87 (42.17)	35.96 (47.55)	-15.60 (76.82)
Student		10.97 (25.89)	15.61 (35.96)	-14.38 (153.19)
Marital Status - Compared to Married				
Never Married		-6.12 (26.27)	-11.96 (29.65)	-12.48 (28.64)
Divorced		-1.93 (31.92)	-8.29 (53.61)	-19.72 (51.83)
Separated		1.66 (19.41)	-27.02 (35.45)	35.97 (67.53)
Widowed		17.43 (58.72)	10.28 (69.92)	9.86 (47.30)
Partners, Not Married		23.16 (33.76)	-58.43 (54.90)	10.28 (59.76)
Political Knowledge - Compared to Very High				
Fairly High			-10.69 (39.88)	-19.22 (48.29)
Average			-11.00 (17.30)	-46.81 (56.67)
Fairly Low			1.52 (19.20)	-33.20 (64.11)
Very Low			-58.43 (54.90)	–

Estimates for parties use Republicans as the reference group
All values rounded to two decimals
Standard Errors shown in parentheses
$^{-}$ No observations
* Significance at the 10% level
** Significance at the 5% level
*** Significance at the 1% level

table 4, there does not appear to be any difference among the demographic groups that are presented here, meaning that this subset's polarization is not well-explained by the demographic characteristics. The most notable results of this exercise, however, are the large point estimates associated with the intercept value and the identification as having "no preference". Because the result of each "party" estimate may be interpreted as the difference between that group and the baseline, we see that there is clearly a difference between the individuals who state no party preference and Republicans, which is perhaps not too surprising. I would expect that truly neutral individuals would not be polarized, and would personally suspect that these individuals have a disinterest in politics which should mean that they are less polarized. Answering "no preference" here is also a bit strange, and would be worth looking into in the future. Since this set is a replication of the set used in Boxell et al. 2017, it includes people who are political independents, but lean to one party or another. The fact that there are some people who claimed to lean towards one party, and later expressed no preference, is at least a little strange. A point estimate as negative as this one provides amusing colloquial evidence that a short memory is depolarizing, but of course this is not a serious point to takeaway from the table as this estimate captures the observation of only a handful of individuals.

To see if there are differences in the way in which the algorithm sorted respondents by treatment effect, I examine the heterogeneity that is observed among subsets which are defined by their estimated treatment effect size. This method closely follows Athey and Wager in defining groups by this estimation and subsequently examining treatment effect differences between them, as was done when looking for heterogeneity based on the mean and modal split for each variable in the causal forest of this thesis. I push that previous analysis forward in this section by also looking at possible projections of the treatment effect among these groups, rather than just between them. The analysis I present proceeds in two steps. First, I use the entire sample and estimate the CATE projections on party identification and, subsequently, demographics within four groups: a "large positive", "small positive", "small negative", and "large negative" group. I define large and small effect sizes in relation to the mean CATE estimate, conditional on being positive or negative. This creates four groups of analysis, respectively defined by $\hat{\tau}_i > \bar{\tau}^+$, $\hat{\tau}_i > \bar{\tau}^+$, $\hat{\tau}_i > \bar{\tau}^-$, and $\hat{\tau}_i < \bar{\tau}^-$, where

$$\bar{\tau}^+ := \frac{\sum\limits_{\{i:\hat{\tau}_i>0\}} \hat{\tau}_i}{|\{i : \hat{\tau}_i > 0\}|} \qquad \text{and} \qquad \bar{\tau}^- := \frac{\sum\limits_{\{i:\hat{\tau}_i<0\}} \hat{\tau}_i}{|\{i : \hat{\tau}_i < 0\}|}$$

Subdividing the groups in this way reveals that party level differences are most pronounced in the groups with a small positive effect and a large negative effect. The intercept for the small positive effect is roughly two points greater than the actual mean point that was used as the threshold (i.e., $\bar{\tau}^+$, which was 5.32). As we would expect, self-declared independents who are within this group are significantly less polarized than their Republican counterparts. Surprisingly, Democrats are significantly less polarized than Republicans at the 10% level once they fall in this group, though with a much more modest point estimate compared to independents. To summarize then, Democratic and Republican respondents who demonstrated modest but positive estimations of treatment were, in general, polarized by the internet- with Republicans becoming more polarized than Democrats- while Independents' level of political disdain was moderated by the introduction of the internet into their homes. The only other group with a significant point estimates in this partition of the data is the "other" category in the regions with large and negative treatment effects. That is to say, among the individuals who the algorithm estimates were the most moderated by the introduction of the internet, respondents who indicated that their political party preference was not for Democrats, Republicans, or Independents showed some of the lowest levels of polarization. With such a strong point estimate as this, I would want to claim that these results could be driven by the fact that these individuals' preferences are distinctly against the traditional party structure, which could mean that these individuals are apathetic about the traditional political organizations in the US. Alternatively, one could look for explanations about how these individuals are so disillusioned with both parties that they rate them at equally (unfavorable) rates. As it turns out, both of these explanations fail for the data examined here. Although some of these individuals were equanimous in their view of both parties, others still show gaps of 70 points between the parties in favor of the Democrats (the party with which they were aligned). Given the large disparities in affect which some exhibited, it seems rather strange that we should see such large and negative point estimates here. However, it ought to be made clear that as we are working with increasingly fine subsections of the population, the statistical results that come from "fringe" answers face serious threats to external validity. By percentage, the "other" responses comprise an already low 1.7 percent of the overall number of respondents categorized as having "large negative responses". In absolute terms, they make up 6.88 weight-adjusted responses out of nearly 400. This means that the significance here is really driven by a handful of individuals, and it would be folly to suppose that their answers extrapolate into the general population in some kind of systematic way. To use the example of one respondent in this category: a treatment effect difference relative to Republicans of -28.05 suggests that, were this individual to become a Republican, the difference in scores that they assign to the parties would be 98.05 points in favor of Republicans[32], just barely over the maximum disparity of 98 points possible given the top-coding of sentiment ratings at 97. Quite simply, more needs to be done here to have the kind of large sample size which is needed to obtain results with external validity. With such a small sample size, there is little that I can justifiably extrapolate from the responses into more general claims about the American electorate.

To close out this section , I use the set of demographic variables which are present in the Ahler et al. 2018 data set (thus, are usable in the next round of analysis) and project it out over two groups of subsets. I first report results for the entire sample separated by the CATE regions from before, and close by using the parties as subsets. Upon examining these results for the first partition, collected in table 6, not many variables are revealed to be significant, however there are some patterns to the variables which are. Among those with a small positive CATE estimations, there are two employment groups who exhibit negative and significant effects: the unemployed and the retired. With around three percent of weighted responses (16.8 total), I would argue that widespread conclusions about the unemployed as a social group are unjustified for reasons similar to the ones I outlined regarding the "other" political category in table 5. The retired make up a more robust 15.2 percent of the category and so have a larger sample with which to work. Compared to the baseline "employed" group, the retired are an estimated 11.57 points less polarized. These effects are also unlikely to be some particular artifact of an age effect which was absorbed by a group of retirees: estimations of these projections with and without employment do not change the fact that the age variable is not significant

[32] Of course, because this individual is a Democratic leaner, I also assume that their 70:0 Democrat-to-Republican sentiment ratio becomes a 0:70. Such an assumption may or may not be realistic depending on how we view the consistency of these possible ratings across parties.

Table 5: Best Linear Projection With Regions Defined By CATE Estimates: Party Identification Projections- Republicans are the Baseline

	Large Positive Effect	Small Positive Effect	Large Negative Effect	Small Negative Effect
	(n = 421)	(n = 562)	(n = 216)	(n = 387)
Intercept	2.29	**7.56****	5.21	-2.62
	(11.79)	(3.82)	(9.62)	(11.01)
Democrat	-4.78	**-4.85***	-2.56	7.66
	(12.64)	(2.58)	(15.61)	(15.48)
Independents	1.65	**-12.26****	-2.95	33.58
	(12.20)	(5.66)	(10.94)	(9.98)
No Preference	-3.74	18.96	13.61	-0.70
	(20.61)	(44.57)	(24.64)	(19.38)
Other	30.30	-21.50	-28.05	0.00
	(44.63)	(25.78)	(11.34)	(51.34)

All values rounded to two digits
Parentheses report standard errors
* Significant at the 10% level
** Significant at the 5% level
*** Significant at the 1% level

here.

The analysis which will most carry over into the Ahler et al. 2018 data set is the projection of treatment effects onto demographic groups, separated by party. Because this analysis which carries over into another specification, and because these results identify treatment differences with a sensible level of "control" (party segmentation) and sufficient observations, I consider the following to be presenting the main results of this thesis. I choose this analysis to carry over because I want to test for effects on out-party stereotyping within that paper's party-group dyads. The effects that are observed to be significant here are the ones which I would expect to also be significant in that set if it were truly the case that stereotyping intensity maps on to the intensity of the partisan's affective gap. Tables 7 and 8 report the results of these tests.

For Democrats, there is a significant and increasing effect along age, implying that we should observe differences in the Ahler et al. 2018 set if we can separate two groups by age. As a way of testing the values of my causal forest, I will segment the groups according the model age split from the forest of Democratic respondents. There is also an effect along ethnic groups for both Democrats and Republicans, in each case between white Americans and the second largest intra-party group. These results are surprising to me because I would not expect there to be a difference in stereotyping along these lines, and so did not expect a difference in polarization. These groups also capture a majority of respondents in both cases, and so are the least likely to be artifacts of the particular survey. As I show with the YouGov stereotyping survey in the next section, there is no underlying difference in stereotyping between any ethnic groups, meaning the gap in polarization between these groups is due to something besides stereotyping.

The most noticeable inter-party difference here is among the educational groups. For Democrats, there are significant differences between the level of affective polarization expressed by the least educated responses and every other level of education individually. Individuals who only finished grade school exhibited a large shift in their affective attachments to the parties due to the introduction of the internet, while the more educated Democrats have estimated effect sizes which imply that they were either slightly polarizedor slightly depolarized by the internet. There is also no such difference among

Table 6: Best Linear Projection Onto Demographic Variables With Regions Defined by CATE Estimates

Term	Large Positive Effect	Small Positive Effect	Large Negative Effect	Small Negative Effect
	($\tau_i \geq 5.37$)	($0 < \tau_i < 5.37$)	($\tau_i \leq -4.29$)	($-4.29 < \tau_i < 0$)
Intercept	-44.08	5.56	7.74	56.72
	(123.59)	(32.17)	(62.29)	(47.34)
Age	0.57	0.09	-0.06	-0.37
	(0.89)	(0.22)	(1.08)	(0.33)
Female	-0.77	5.33	-13.10	**-11.40**$^{[*]}$
	(6.55)	(5.63)	(21.48)	(6.63)
Ethnicity - Compared to Whites				
Black	-1.85	8.56	-5.86	5.02
	(32.29)	(7.05)	(25.61)	(8.81)
Hispanic	25.60	1.56	-2.88	16.32
	(20.53)	(2.07)	(24.99)	(11.39)
Other or Multiple Races	13.70	-3.00	**-36.51*****	25.53
	(83.86)	(10.20)	(6.63)	(27.02)
Education - Compared to Grade School or Less				
High School	1.48	-4.15	27.59	-29.73
	(62.71)	(9.27)	(26.35)	(35.38)
Some College	9.23	4.66	34.29	-31.51
	(78.40)	(16.91)	(40.69)	(37.88)
College or Advanced Degree	13.70	3.41	28.36	-38.40
	(83.68)	(14.95)	(43.26)	(42.18)
Income - Compared to the Lowest 16%				
17th to 33rd Percentile	9.27	-13.13	-0.27	-13.91
	(24.05)	(38.32)	(25.06)	(15.23)
34th to 67th Percentile	3.97	-10.84	**-26.63****	-2.02
	(36.61)	(28.87)	(11.95)	(12.67)
68th to 95th Percentile	7.08	-11.88	**-31.68*****	2.72
	(30.70)	(25.65)	(11.08)	(20.50)
96th to 100th Percentile	12.42	-13.94	-26.84	8.86
	(19.70)	(27.96)	(31.97)	(28.19)
Employment Status - Compared to Currently Working				
Temporarily Laid Off	13.39	-10.24	31.24	12.79
	(20.03)	(30.50)	(21.08)	(39.18)
Unemployed	7.84	**-16.09*****	-8.75	31.84
	(42.78)	(5.40)	(12.94)	(23.46)
Retired	2.43	**-11.57****	23.73	12.15
	(18.77)	(5.00)	(36.06)	(14.35)
Permanently Disabled	27.21	-5.46	3.90	5.74
	(51.97)	(17.63)	(30.76)	(22.68)
Homemaker	18.26	7.71	1.12	-16.42
	(28.34)	(5.75)	(18.94)	(32.37)
Student	16.76	-3.38	10.58	-2.32
	(56.00)	(7.49)	(37.84)	(38.50)
Marital Status - Compared to Married				
Never Married	6.94	1.25	19.59	-4.78
	(15.04)	(4.56)	(30.91)	(7.88)
Divorced	-18.56	-5.67	-4.49	1.22
	(34.50)	(7.32)	(27.53)	(9.68)
Separated	-1.52	5.59	**-37.02**$^{[**]}$	-7.73
	(24.86)	(26.27)	(16.40)	(19.89)
Widowed	2.21	12.25	-42.56	-14.90
	(24.78)	(9.95)	(32.17)	(9.51)
Partners, Not Married	15.61	4.42	-10.15	-4.96
	(22.12)	(9.36)	(41.42)	(27.03)

All values rounded to 2 decimal places

Parentheses report standard errors

* Significance at the 10% level

** Significance at the 5% level

*** Significance at the 1% level

Table 7: Linear Projections from Demographic Variables - Democrats

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Intercept	-6.28	1.91	-0.51	**39.46****	13.24	0.90	**1.90***	**33.18***
	(4.25)	(3.11)	(2.61)	(16.76)	(9.83)	(1.14)	(1.09)	(19.30)
Age	**0.15***							0.30
	(0.08)							(0.33)
Female		-2.34						0.36
		(5.67)						(8.48)
Ethnicity - Compared to Whites								
Blacks			**-9.85***					-18.25
			(5.02)					(23.28)
Asian or Pacific Islander			-0.001					-0.13
			(30.59)					(34.24)
American Indian or Alaska Native			-9.09					-11.39
			(56.65)					(69.11)
Hispanic			20.10					18.66
			(15.07)					(16.09)
Other or Multiple Races			15.82					**17.26****
			(12.80)					(8.72)
Education- Compared to Grade School								
High School				**-41.09***				-31.31
				(21.83)				(20.95)
Some College				**-38.16****				**-28.14***
				(17.28)				(16.77)
College or Advanced Degree				**-37.97***				-29.14
				(17.02)				(17.96)
Family Income - Compared to Bottom 16%								
17 to 33 Percentile					-9.32			-12.20
					(15.82)			(22.69)
34 to 67 Percentile					**-17.26****			-20.52
					(7.47)			(15.32)
68 to 95 Percentile					-12.52			-15.98
					(9.52)			(15.47)
96 to 100 Percentile					-7.18			-11.40
					(17.19)			(20.34)
Work Status- Compared to the Employed								
Temporarily Laid Off						11.76		7.31
						(25.00)		(14.60)
Unemployed						7.07		11.53
						(19.73)		(21.05)
Retired						1.76		-3.64
						(6.88)		(12.49)
Permanently Disabled						7.24		5.92
						(12.91)		(13.72)
Homemaker						-8.38		-10.54
						(8.20)		(13.39)
Student						-6.07		-12.05
						(11.17)		(16.59)
Marital Status - Compared to Married								
Never Married							4.16	3.48
							(5.43)	(9.84)
Divorced							-18.60	-23.23
							(19.66)	(20.87)
Separated							23.75	22.29
							(21.29)	(18.45)
Widowed							-6.61	-16.54
							(8.09)	(23.81)
Partners; Not Married							0.18	2.73
							(17.11)	(18.32)

Parenthetical Values Report Standard Errors
All Values Rounded to 2 digits
* Significance at the 10% level
** Significance at the 5% level
*** Significance at the 1% level

Table 8: Linear Projections from Demographic Variables - Republicans

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Intercept	**6.66****	3.81	2.38	-20.18	**5.56*****	0.82	2.48	-6.35
	(3.33)	(3.78)	(2.06)	(19.17)	(1.59)	(1.30)	(1.97)	(19.29)
Age	-0.09							**-0.23****
	(0.09)							(0.11)
Female		-2.55						-5.09
		(5.50)						(6.23)
Ethnicity - Compared to Whites								
Blacks			1.14					3.02
			(10.68)					(11.79)
Asian or Pacific Islander			**-14.73*****					**-16.72*****
			(2.89)					(5.13)
American Indian or Alaska Native			-7.37					-7.87
			(13.90)					(10.48)
Hispanic			1.26					1.05
			(6.87)					(7.62)
Other or Multiple Races			6.36					7.63
			(23.60)					(24.99)
Education- Compared to Grade School								
High School				20.61				18.09
				(18.58)				(15.67)
Some College				25.88				24.77
				(20.46)				(17.33)
College or Advanced Degree				23.10				24.07
				(18.66)				(16.76)
Family Income - Compared to Bottom 16%								
17 to 33 Percentile					-5.63			**-6.91****
					(6.05)			(3.48)
34 to 67 Percentile					-1.76			-3.00
					(5.43)			(3.10)
68 to 95 Percentile					-3.56			-4.55
					(6.75)			(8.83)
96 to 100 Percentile					**-10.10*****			**-11.58***
					(2.31)			(6.34)
Work Status- Compared to the Employed								
Temporarily Laid Off						7.92		10.07
						(9.51)		(11.37)
Unemployed						-11.63		-14.02
						(12.35)		(14.45)
Retired						5.43		**14.85***
						(7.08)		(7.73)
Permanently Disabled						3.41		4.67
						(15.77)		(16.99)
Homemaker						16.90		21.46
						(10.97)		(13.98)
Student						5.51		-0.18
						(15.10)		(13.78)
Marital Status - Compared to Married								
Never Married							3.53	2.93
							(4.10)	(3.29)
Divorced							1.82	3.19
							(7.56)	(7.43)
Separated							**-18.04***	**-17.20*****
							(10.41)	(6.10)
Widowed							-4.42	-4.37
							(8.58)	(9.47)
Partners; Not Married							0.32	-1.06
							(25.49)	(28.96)

Parenthetical Values Report Standard Errors
All Values Rounded to 2 digits
* Significance at the 10% level
** Significance at the 5% level
*** Significance at the 1% level

the group of Republicans along similar lines. I reason that there are two potential explanations behind these sets of effects. It could be that the lowest educated Democrats are a particularly polarized group, which creates the large and significant intercept value and the sharp contrasts. Instead, it could be that the lowest educated Democrats and Republicans are polarized and that there are not differences along education within the Republican party. To test these ideas, I project the treatment effects of the lowest educated onto the respondents' party leanings. I find that the lowest educated Republicans in this group had significantly higher levels of polarization at the 10 percent level, with a point estimate of 59.63 points. This would mean that nonsignificant differences among Republicans points to a higher base level of polarization among all education groups within the party. As I will also discuss in section 4.3, I believe that these effects uncover the mechanism of polarization apparent within my data. When individuals are assessing the parties, they are doing so by evaluating party-prototypical groups. A more diverse party is more difficult to judge in this way because there are more intra-party groups to assess, and so respondents would need to alter their assessments to attempt to account for these different groups. Because the Republican party is less diverse than the Democratic party (which one can see by looking at the demographic breakdowns in table 1), it is easier to stereotype. I believe that education effects also moderate this behavior, perhaps through differences in formal critical thinking education, and result in a diverse party which is less prone to stereotyping as one moves along the education scale, and a less diverse one which can be grouped together (erroneously) more easily.

4.3 Measurements of Stereotyping

To close the analysis of this thesis, I consider previous work by Ahler and Sood, primarily Ahler et al. 2018. This paper finds that there is, in general, a relationship between partisan affect and stereotyping, and that there are certain explanatory mechanisms which can be ruled out- notably innumeracy, ignorance of base rates, and expressive responding. That is to say, respondents in that study were frequently incorrect about the extent to which the parties contained "prototypical partisans", and this misperception was not due to having population percentages in their heads which summed to a number greater than 100, nor was it because they failed to understand how numerous a social group (e.g. the lesbian, gay, and bisexual community) was, nor was it because their answers were intentionally inflated to make a point. This tells us ex-ante that the type of heuristic pitfalls into which our respondents fall are almost certainly due to representativeness, which links back to the equation for distortions in probability due to stereotyping that is articulated in Gennaioli et al. 2019, and would be exacerbated by the introduction of the internet. For my part then, I seek to extend this paper in two important ways: first, I provide cursory evidence that the mechanism underpinning partisan animus due to stereotyping is the result of prejudice rather than attribution of extreme positions to individuals in the party-prototypical social group, at least among Republicans. Second, I examine some of the data used in Ahler et al. 2018 to look at heterogeneity in stereotyping across social groups as identified in ANES, seeking to find whether individuals with views that showed heavy heterogeneity here are also those with highly distorted views in that work.

Ahler et al. 2018 employs attitudes about four party-stereotypical groups per party, constructing eight party-group dyads which provide the basis of their estimation of distortions due to stereotyping. For Democrats, these groups are Americans who are black; labor unions members; lesbian, gay, and bisexual Americans; and atheists and/or agnostics. The groups for Republicans are evangelicals, southerners, those over age 65, and those who make more than $250,000 per year. Because I want to understand how the treatment of the internet differed among respondents, and to project those differences onto variables which capture opinions about these groups, I employ ANES's thermometer ratings. However, limitations within ANES data allow me to only estimate effects for a fraction of the party-social group dyads which are used in that paper, and often with imperfect mapping between those questions and the ones ANES asks. I thus report results concerning projections onto measures of warmth towards Americans who are black, Americans who are lesbian or gay, and labor unions. Note that each of these groups is party-prototypical for Democrats, but that the measures may be imperfect. Of these, the poorest measure is that on labor unions, which is not asking about labor union members as in Ahler et al. 2018, a caveat which needs to be mentioned for interpreting results. As to the lack of Republican-prototypical group sentiments: variable VCF0208 in the cumulative ANES data file does ask about Southerners, but the measure only goes up to 2008 and, thus, cannot enter into this exercise because it wasn't a common variable for the entire time series and, as such, was dropped. Similarly, VCF9268 asks about feelings towards the rich, but was introduced in 1998; VCF9004 asks about the elderly but was phased out in 2004; and VCF9003 asks about evangelicals but was phased out in 1988. Due to this combination of unfortunate timing in the phasing out of questions, I can only report on the results concerning Republicans' feelings towards the Democratic groups.

Table 9(a) shows the results of best linear projections of the treatment effect from causal forest 4 (Republicans on Democrats), applied to the three party-stereotypical groups which I can recover from ANES data. The "full" model, i.e. the projection onto all three variables simultaneously, demonstrates that the base treatment effect which the algorithm obtains is quite large, indicating that a Republican who felt extremely negatively about all three groups would have experienced a large shock in their level of animosity towards Democrats upon the introduction of the internet. However, none of the variables themselves have significant point estimates. This result makes follows the logic of Ahler et al. 2018: were individuals to have a picture of the parties in their heads, then featuring all three stereotypical groups in one estimation would be akin to estimating a model with a large degree of multicollinearity. For this reason, my preferred specifications are the first three, which just examine the bivariate relationships between treatment effect estimate and sentiments towards a single group[33].

Here, we see that Republican sentiment towards Americans who are black has bearing on their treatment effect, which we would expect given the high degree of Democratic party (mis)perception which is observed in Ahler et al. 2018. The nonsignificant results would seem to run against the findings of that paper, but there are reasons to believe that these are simply artifacts of the data. That attitude towards labor unions is not a significant variable is not too surprising for the reasons mentioned above about the difference between the unions as entities and the members as individuals. For the lesbian and gay community, there are two different factors interfering with proper estimation. The more important of these, as mentioned before, is the imperfect overlap between the Ahler et al. 2018 question and this one, which does not include the bisexual community. The second is the rapid change in Americans' general sentiment towards the community over this time span. In 1996, the average score on this rating was 39.27. By 2000, it had risen to 46.7, and by 2012 it was 53.46. I suspect that this projection has a diffi-

[33] This also more closely mirrors the examination of party-group dyads which is used in Ahler et al. 2018

Table 9: Exploration of Projections Informed by Stereotypes, for Republicans

(a) Best Linear Projections of Treatment Effects on Republicans to Attitudes About Democratic-Prototypical Groups

	(1)	(2)	(3)	(4)
Intercept	**-25.29**** (11.24)	-7.78 (5.32)	-7.45 (11.59)	**-26.22***** (8.4)
Thermometer- Black Americans	**0.32**** (0.16)			0.31 (0.26)
Thermometer- Lesbians and Gays		0.088 (0.09)		0.01 (0.25)
Thermometer- Labor Unions			0.08 (0.23)	0.04 (0.11)

(b) Treatment Effect Heterogeneity Projected onto Attitudes About Black Americans and The Liberal-To-Conservative Scale for Democrats

	All	All	"Low" Affect	"High" Affect
Intercept	-8.94 (9.21)	**-31.78***** (9.32)	-12.51 (9.47)	-6.44 (9.27)
Democrats: Liberal-to-Conservative	2.25 (3.58)	2.63 (3.42)	1.55 (4.44)	3.26 (3.12)
Thermometer- Black Americans		**0.034**** (0.15)		

Parenthetical Values Report Standard Errors
* Significance at the 10% level
** Significance at the 5% level
*** Significance at the 1% level

cult time distinguishing between the treatment effects which were uniquely due to the internet with the secular trends which were taking place during this period of time, resulting in larger than necessary standard error estimates due to the more variable nature of this measure. Table 9(b) provides some evidence that the mechanism underpinning the increased animosity has more to do with prejudice than with attribution of extreme positions to out-party prototypical groups, answering a lingering question of Ahler et al. 2018. A combined model which features both the sentiment towards Americans who are black and the respondent's opinion of the Democratic party on the liberal-to-conservative scale reveals significant effects for the former but not the latter. If it were the case that the misperception of the numerosity of Americans who are black in the Democratic party drove partisanship through an assumption that Americans who are black held more extreme positions which were unpalatable to the respondent, then we would expect that the treatment effect would be heterogeneous along the variable wherein the respondent defines just how "liberal" they believe the party to be. However, this is not the case when we combine these measures together. Only the sentiment towards Americans who are black is significant. Moreover, in checking this result by splitting the respondents into a high and low group based on their affect towards Americans who are black[34], similar to what I did in the last section, I still do not detect heterogeneity in this variable between either of these groups. I also don't detect significant effects in the full sample with sentiment towards Americans who are black removed. To sum then, there is stronger evidence that

[34] In truth, very few people answer below 50 on this thermometer question, so to control for the potential distortion in this measure due to social desirability, I define high and low groups relative to the conditional mean sentiment above 50.

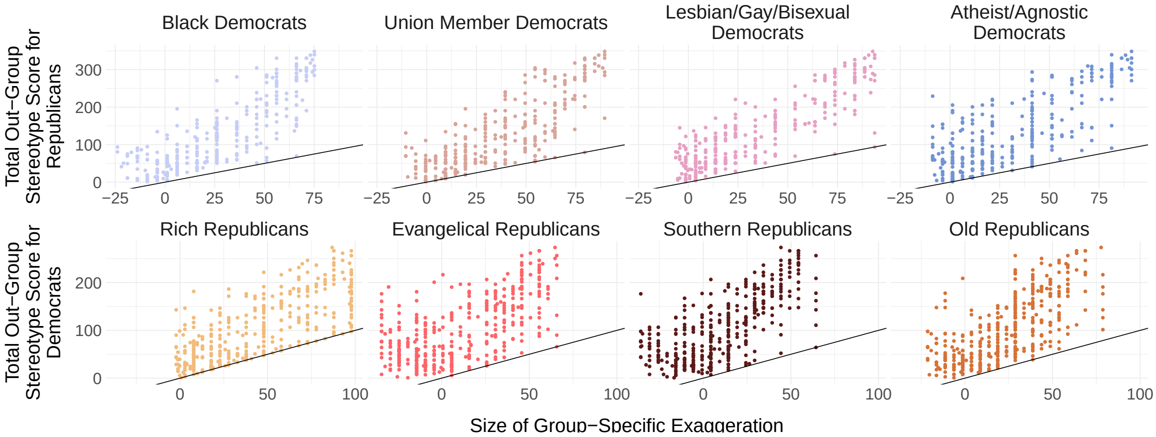


Figure 3: Size of specific exaggerations against total stereotype score. The 45 degree line shows the lowest possible score, corresponding to a positive exaggeration on the specific score and nonpositive scores for the rest. Negative exaggerations mean that the respondent underestimated the numerosity of the group in the party

the effect which the internet had on polarization operated through a prejudicial mechanism than there is evidence that it operated through an association with extreme partisan positions.

To adapt the YouGov data from Ahler et al. 2018 for my purposes, I perform a few changes. First, I remove answers from the survey where a respondent's answer is quite clearly nonserious, by which I mean answering 100 or 0 for every party-group dyad for at least one party[35]. Second, I derive a stereotype score for each respondent which is the sum of all responses for the party group dyads of the other party in which their estimate of the numerosity of a group is in excess of the true percentage. Figure 3 shows the scatter plots for group-specific exaggerations against the total measure for both parties. I choose this as my stereotype score because I want to show differences in distorted perceptions of the representativeness of groups in a way which does not assume under-guessing to be the same as over-guessing (hence no absolute values) or which cancels out over-guessing with under-guessing (hence no pure "columnwise" sums).

Further, table 10 reports the results of two sample (Welch's) t-tests for differences in mean stereotype scores between groups in the YouGov data from Ahler et al. 2018. I choose to report results either when a test for differences is justified by a significant difference in tables 7 or 8, in which case the alternative hypothesis follows the suggested direction of differences from those results, or when a group comparison that was not significant in the previous tables is significant in this data set, in which case the alternative hypothesis is two-tailed because I have no reason to suppose the direction of difference ex-ante. In the case of age, which does not have a baseline comparison group, I split the variable according to the modal value suggested by the causal forest for that party[36]. Because of limited sample size, I am unable to provide testing from the final column of each table.

An initial scan of this table reveals that the two data sets depart from one another

[35] I feel that this group is the only one which I can really remove on some justifiable grounds. I don't regard a respondent who states, for example, that every member of the Democratic party is a black atheist union-member who is part of the LGB community to be giving a serious answer.

[36] That is to say, 37 is the value which is used for Democrats because 37 was the modal split value in the causal forest for Democratic respondents on Republicans

Table 10: Results of Two-Sample t-tests for Differences in Means of Stereotyping Score

Party	Group 1	Group 2	Test Direction	Mean 1	Mean 2	t-statistic
Democrats	Younger than 37	37 or Older	Less	95.51	105.26	**-1.43*** (0.08)
By Ethnicity						
Democrats	Whites	Blacks	Greater	96.73	108.60	-1.37 (0.91)
Republicans	Whites	Asian or Pacific Islander	Greater	96.73	72.54	1.20 (0.26)
By Education						
Democrats	Grade School	High School	Greater	123.87	106.27	1.30 (.10)
Democrats	Grade School	Some College	Greater	123.88	88.94	**2.64***** (0.01)
Democrats	Grade School	College or Advanced Degree	Greater	123.88	98.36	**1.88**** (0.03)
By Income Percentile						
Democrats	Bottom 16th	34th to 67th	Greater	89.33	92.91	-0.38 (0.65)
Democrats	Bottom 16th	68th to 95th	Two-Sided	89.33	114.34	**-2.41**** (0.02)
Republicans	Bottom 16th	Top 4%	Greater	117.07	101.45	0.45 (0.33)

All estimates rounded to two digits
Parentheses report p values for the corresponding test statistic
All standard errors are calculated via bootstrapping
Under Direction: "Greater" signifies that the test is one-sided with Mean 1 > Mean 2 as the alternative hypothesis. "Less" signifies the opposite alternative hypothesis. One-sided tests are only used if differences were significant in the ANES set (see tables 7 and 8)
All tests carried out with "weights" package in R.
* Significant at the 10% Level
** Significant at the 5% Level
*** Significant at the 1% Level

with respect to the groups that have significant differences. However, the general spirit of the bifurcations is largely preserved. For example, there is clearly a difference in stereotyping behavior among the dimension of schooling for Democrats, with the lowest educated performing worse than their fellow Democrats, as was shown in the ANES data set. This point is worth lingering on as it is the most directly related to the way I would imagine respondents deal with questions about stereotypical groups. As posited in Ahler et al. 2018, this is evidence that, when individuals are asked to think about the parties, they think about them in terms of other groups which are more readily available in their mind, such as race and class. I would expect that those individuals subsequently temper their views about the party in some attempt to divorce group association with party association, but that this skill differs by education. Hence, there is nearly perfect overlap between the significant (and nonsignificant) results of the ANES data and this one for both Democrats and Republicans, with the only difference being the comparison between the least educated and second-least educated Democrats[37]. The great surprise here might be that these education effects are not equal for both parties, which could either be due to a particularly egregious tendency towards polarization on the part of lower educated Democrats compared to their peers, or a sort of equivalence among Republicans of all education levels. A priori, there's no reason to suppose one over the other, though I suspect that this result relates to the similar findings from table 7. The general tendency of the Democratic party to be more diverse could imply that any individual would have a difficult time doing the necessary unscrambling that is required to have a clear picture of the party from messy, stereotypical pictures. If this general line of reasoning is correct, it would also mean that obtaining a corrected idea of the composition of the Republican party is easier, and one would expect it to be easier as an increasing function of education. A two-tailed test for the differences of means between the lowest-educated Democrats and Republicans has insignificant effects with a p-value of nearly 0.5, indicating that the differential effects by education between the parties are not driven by an extreme difference in baselines. Not only do these results demonstrate that the differences in polarization brought on by the internet correlate with differences in stereotyping out-groups more generally, they suggest a potential pathway whereby intra- and inter-party differences emerge. It is also quite apparent that the differences between Democrats who are white and those who are black were not caused by the underlying differences in stereotypes because the test would only be significant were we to be testing in the opposite direction of that implied by the ANES differences.

5 Discussion

This thesis is, to my knowledge, the first to apply generalized random forests to the American National Election Survey. As was made apparent in the section on methodology, there are still grf package developments currently in the works which would make these results more compelling. The most desirable feature would be a strategy to deal with missing values, such that researchers in the future need not remove large portions of the data before analysis has begun. Having more data would allow for finer tuning of the forests than I was able to accomplish here, and more reliable cluster estimation. More data would also allow future work to test the limits of the software's capabilities, making

[37] And even in that case, the p-value for the associated test 0.10034, just shy of significance at the 10% level

intentionally conservative decisions about tuning parameters that make detecting effects as difficult as is reasonable, so as to improve the credibility of results. I would also be interested in understanding the differences in polarization along longer time-spans, which was not feasible in this context due to my use of the internet as a treatment. Future work in this space could more closely examine year-by-year effects along a common, and maybe more parsimonious, set of variables which build off of the important measures discovered here. Nonetheless, it is clear that the use of generalized random forests is promising for researchers who wish to understand phenomena with complex data generating processes and, in particular, who do not wish to assume linearity of response functions or Gaussianity, but still want to obtain valid confidence intervals and perform statistical tests.

I find empirical evidence that political polarization and stereotyping are related phenomena which tend to occur along the same social divisions. Subdivisions of the population reveal that the estimated treatment effects provide decent subsetting criteria by which to understand different slices of the population, and such an examination in this setting revealed a general difference between Democrats, Republicans, and Independents in some of those groups (namely, among those for whom the internet had the effect of moderately increasing affective polarization). However, the performance of these algorithms when placing observations into subsets according to their CATE estimations and projecting onto demographic variables showed dubious effects at best, which may be the result of the low sample size available. My estimates suggest that there are not differences in polarization along "purely" demographic lines (as opposed to social strata lines such as income) in the general population, but there may be within the parties. In particular, there is no significant difference in polarization between ethnic groups, between ages, or between genders. These results are not too surprising given the staggering amount of research before this thesis which has looked into the problem of political polarization; indeed, it would be quite concerning for evaluating the performance of my algorithms if I were to find effects. Instead, there is compelling evidence that there are profound connections between polarization and education. The driving force behind this connection is still unknown, but I strongly suspect that differences in heuristic thinking ought to be considered as the culprit. This is particularly true given that the main intra-party differences in propensity towards both affective polarization and stereotyping are education based, suggesting that training in critical thinking skills has the potential to help overcome such biases. Moreover, the complicated interactions of multiple identities would imply that it is much easier to stereotype about Republicans than about Democrats, who are comparatively more diverse. Education effects subsequently counteract leanings towards stereotyping by moderating overblown assumptions about out-party composition and getting closer to the true value (signifying the engagement in "slow thinking" as Kahneman 2013 puts it). Since there is no inter-party difference between the two baseline, lowest educated groups in their propensity to stereotype, this is my preferred explanation[38]. However, more would need to be done in future work in order to actually assess differences in propensities towards cognitive biases, ideally through survey design which includes a battery of demographic and political viewpoints.

The American journalist Ezra Klein, in his recent book Why We're Polarized, writes about polarization as a normal phenomenon which has been interrupted by anomalous

[38] This can also be seen in the way that polarized party members talk about the other side in online discourse. Republicans are stereotyped according to a race, an education level, and a religion (e.g. "white Christian rednecks") while Democrats are often stereotyped according to a lifestyle, and occasionally a lack of religion (e.g. "latte-sipping atheists").

non-polarization during the 20th century. One of the central theses of the book is that the polarization which is observed in modern American politics is the result of a complicated system of incentives that encourage rational actors to polarize themselves in order to hold power. Once the parties sorted themselves after the passage of the Civil Rights Act and the "Dixiecrats" were dislodged from the left (also documented in Mason 2013), it became ever-easier for politicians within the system to consolidate intra-party power, and maximize inter-party contrast. Setting aside the issues around the supply of politicians for other works, this thesis corroborates the notion that sorting begets affective polarization; political parties, once well sorted, become easier to (affectively) polarize around the common, and more primitive, markers of identity which increasingly map onto the parties. These markers operate on some of the deepest cognitive considerations that humans encounter every day: the need to feel as part of a broader group, the desire to see that group succeed, the tendency to punish the members of the out-group (Tajfel et al. 1974), and, most importantly for this work, the inclination to see the out-group as a caricatured version of itself. The problem of political polarization is likely to only get worse as sorting behavior encourages "mega-identities" to stack upon one another, but my findings suggest that it is possible to roll the problem back, if only the parties were able to engage in the slow thinking necessary to better see one another as they are, and not as they are stereotyped to be.

References

Abramowitz, A. I. and K. L. Saunders (2008). "Is Polarization a Myth?" In: *The Journal of Politics* 70.2, pp. 542–555.

Ahler, D. J. and G. Sood (2018). "The Parties in Our Heads: Misperceptions about Party Composition and Their Consequences". In: *The Journal of Politics* 80.3, pp. 964–981.

Alesina, A., S. Stantcheva, and E. Teso (2018). "Intergenerational Mobility and Preferences for Redistribution". In: *American Economic Review* 108.2, pp. 521–554.

Athey, S. and G. Imbens (2016). "Recursive partitioning for heterogeneous causal effects". In: *Proceedings of the National Academy of Sciences of the United States of America* 113.27, pp. 7353–7360.

Athey, S., J. Tibshirani, and S. Wager (). *Generalized Random Forests.*

Athey, S. and S. Wager (). *Estimating Treatment Effects with Causal Forests: An Application.*

Atkin, D., E. Colson-Sihra, and M. Shayo (2019). *How Do We Choose Our Identity? A Revealed Preference Approach Using Food Consumption.* Cambridge, MA.

Biau, G. and E. Scornet (2016). "A random forest guided tour". In: *TEST* 25.2, pp. 197–227.

Bill Bishop (2017). *The Big Sort: Why the Clustering of Like-minded America Is Tearing Us Apart.* Tantor Media Inc.

Bordalo, P., K. Coffman, N. Gennaioli, and A. Shleifer (2016). "Stereotypes". In: *The Quarterly Journal of Economics* 131.4, pp. 1753–1794.

Boxell, L., M. Gentzkow, and J. M. Shapiro (2017). "Is the Internet Causing Political Polarization? Evidence from Demographics". In: *NBER Working Papers Series.*

— (2019). *Cross-Country Trends in Affective Polarization.*

Breiman, L., J. Friedman, R. Olshen, and C. Stone (1984). *Classification and Regression Trees.* new edition **cart93**? Monterey, CA: Wadsworth and Brooks.

Breiman, L. (2001). "Random Forests". In: *Machine Learning* 45, pp. 5–32.

Chernozhukov, V., M. Demirer, E. Duflo, and I. Fernández-Val (). *Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments.*

Chipman, H. A., E. I. George, and R. E. McCulloch (2010). "BART: Bayesian additive regression trees". In: *The Annals of Applied Statistics* 4.1, pp. 266–298.

Fiorina, M. P., S. J. Abrams, and J. Pope (2011). *Culture war? The myth of a polarized America.* 3. ed. Great questions in politics series. Boston, Mass.: Longman.

Fleisher, R. and J. R. Bond (2004). "The Shrinking Middle in the US Congress". In: *British Journal of Political Science* 34.3, pp. 429–451.

Gennaioli, N. and G. Tabellini (2019). "Identity, Beliefs, and Political Conflict". In: *CESifo Working Paper.*

Huber, G. A. and N. Malhotra (2017). "Political Homophily in Social Relationships: Evidence from Online Dating Behavior". In: *The Journal of Politics* 79.1, pp. 269–283.

Iyengar, S., Y. Lelkes, M. Levendusky, N. Malhotra, and S. J. Westwood (2019). "The Origins and Consequences of Affective Polarization in the United States". In: *Annual Review of Political Science* 22.1, pp. 129–146.

Iyengar, S. and S. J. Westwood (2015). "Fear and Loathing across Party Lines: New Evidence on Group Polarization". In: *American Journal of Political Science* 59.3, pp. 690–707.

Kahan, D. M. (2015). "Climate-Science Communication and the Measurement Problem". In: *Political Psychology* 36.1, pp. 1–43.

Kahneman, D. (2013). *Thinking, fast and slow.* 1. paperback ed. Psychology/economics. New York: Farrar Straus and Giroux.

Klein, E. (2020). *Why we're polarized.* First Avid Reader Press hardcover edition. New York: Avid Reader Press.

Mason, L. (2013). "The Rise of Uncivil Agreement". In: *American Behavioral Scientist* 57.1, pp. 140–159.

Rubin, D. B. (2005). "Causal Inference Using Potential Outcomes". In: *Journal of the American Statistical Association* 100.469, pp. 322–331.

Semenova, V. and V. Chernozhukov (). *Estimation and Inference about Conditional Average Treatment Effect and Other Structural Functions.*

Shayo, M. (2009). "A Model of Social Identity with an Application to Political Economy: Nation, Class, and Redistribution". In: *American Political Science Review* 103.2, pp. 147–174.

Su, X., C.-L. Tsai, H. Wang, D. M. Nickerson, and B. Li (2009). "Subgroup Analysis Via Recursive Paritioning". In: *Journal of Machine Learning Research* 10, pp. 141–158.

Sunstein, C. R. (2002). "The Law of Group Polarization". In: *Journal of Political Philosophy* 10.2, pp. 175–195.

Tabellini, G. (2010). "Culture and Institutions: Economic Development in the Regions of Europe". In: *Journal of the European Economic Association* 8.4, pp. 677–716.

Tajfel, H. and J. C. Turner (1974). "The Social Identity Theory of Intergroup Behavior". In: *Political Psychology.* Ed. by J. T. Jost and J. Sidanius. Psychology Press, pp. 276–293.

Tibshirani, J., S. Athey, and S. Wager (2019). *grf: Generalized Random Forests.* R package version 1.0.1.

Wager, S. and S. Athey (2018). "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests". In: *Journal of the American Statistical Association* 113.523, pp. 1228–1242.