Methodological Appendix for "Assessing the Russian Internet Research Agency's Impact on the Political Attitudes and Behaviors of Twitter Users in the United States in Late 2017"

Contents

Introduction

This methodological appendix provides a detailed description of the data and methods described in Bail et al.'s (2019) study "Assessing the Impact of the Russian Internet Research Agency on the Political Attitudes and Behaviors of American Twitter Users in Late 2017." Upon publication of this manuscript, all code and data will be made available via this link on the Dataverse. Some of the text and figures presented below originally appeared in Bail et al.'s earlier (2018) study, from which the survey data for the current study were provided.

Survey Recruitment Process

Bail et al. (2018) hired YouGov—one of the largest survey firms in the United States—to recruit at least 1,200 self-identified Republicans and Democrats over age 17 who visit Twitter at least three times each week to complete five surveys between mid-October 2017 and mid-November 2017. A more detailed description of YouGov's online nonprobability panel is available here.

Figure 1 (below) provides a detailed description of the recruitment process for the first survey which was fielded between October 10th and October 19th, 2017. YouGov invited 10,634 members of its U.S. panel to participate in the study using U.S. census sampling frames. Of these, 5,520 did not respond,

www.pnas.org/cgi/doi/10.1073/pnas.1906420116

and 5,114 accepted the invitation, for an initial cooperation rate of 48%. These individuals were then asked several screening questions. First, they were asked about their party identification using the following question: "Generally speaking, do you think of yourself as a [Democrat/Republican/Independent/Other/Not Sure]." Respondents who did not respond with either "Democrat" or "Republican" were screened out, and remaining respondents were stratified according to their answer to the following question in order to balance the sample according to strength of partisanship: "Would you call yourself a strong Democrat/Republican or a not very strong Democrat/Republican?" These two questions have been widely employed to measure party attachment in the American National Election Study and many other surveys. Third, respondents were asked if they "visit Twitter at least three times a week in order to read messages from other Twitter accounts," and screened out if they answered negatively.

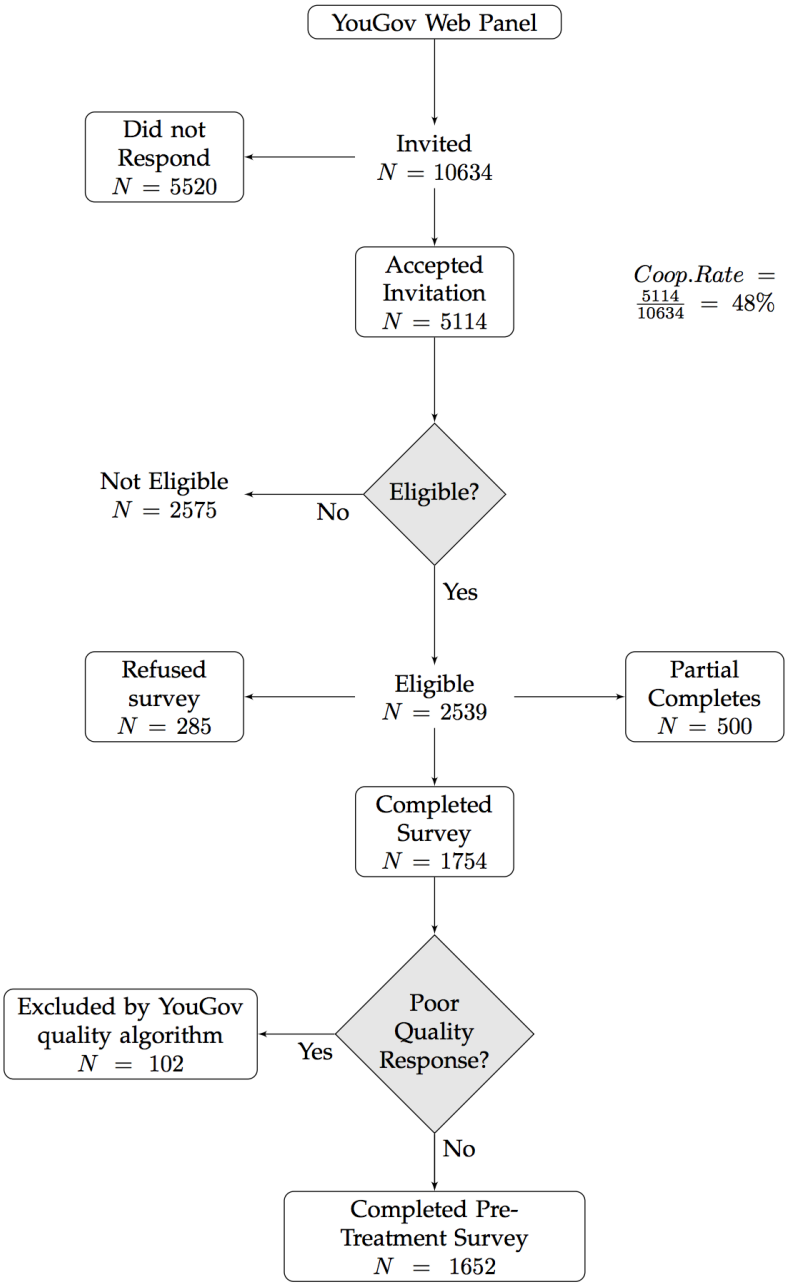

Figure S 1: Recruitment Process for First Survey, Reprinted from Bail et al. (2018)

A total of 2,539 people were deemed eligible according to these two initial eligibility criteria, and were subsequently re-directed to an informed consent dialogue and offered the equivalent of $11 via YouGov's "points" system—which allows respondents to redeem points for items such as Amazon gift cards—to share their Twitter handle, or Twitter ID, in order that it may be linked to their survey responses.

In the informed consent dialogue pictured in Figure S2 (below), eligible pre-treatment survey participants were informed that the survey would take about 10 minutes and was designed to "investigate peoples' experiences on Twitter." They were also informed that participants who completed this survey and provided a valid Twitter handle would be eligible to complete a follow-up survey one month later.

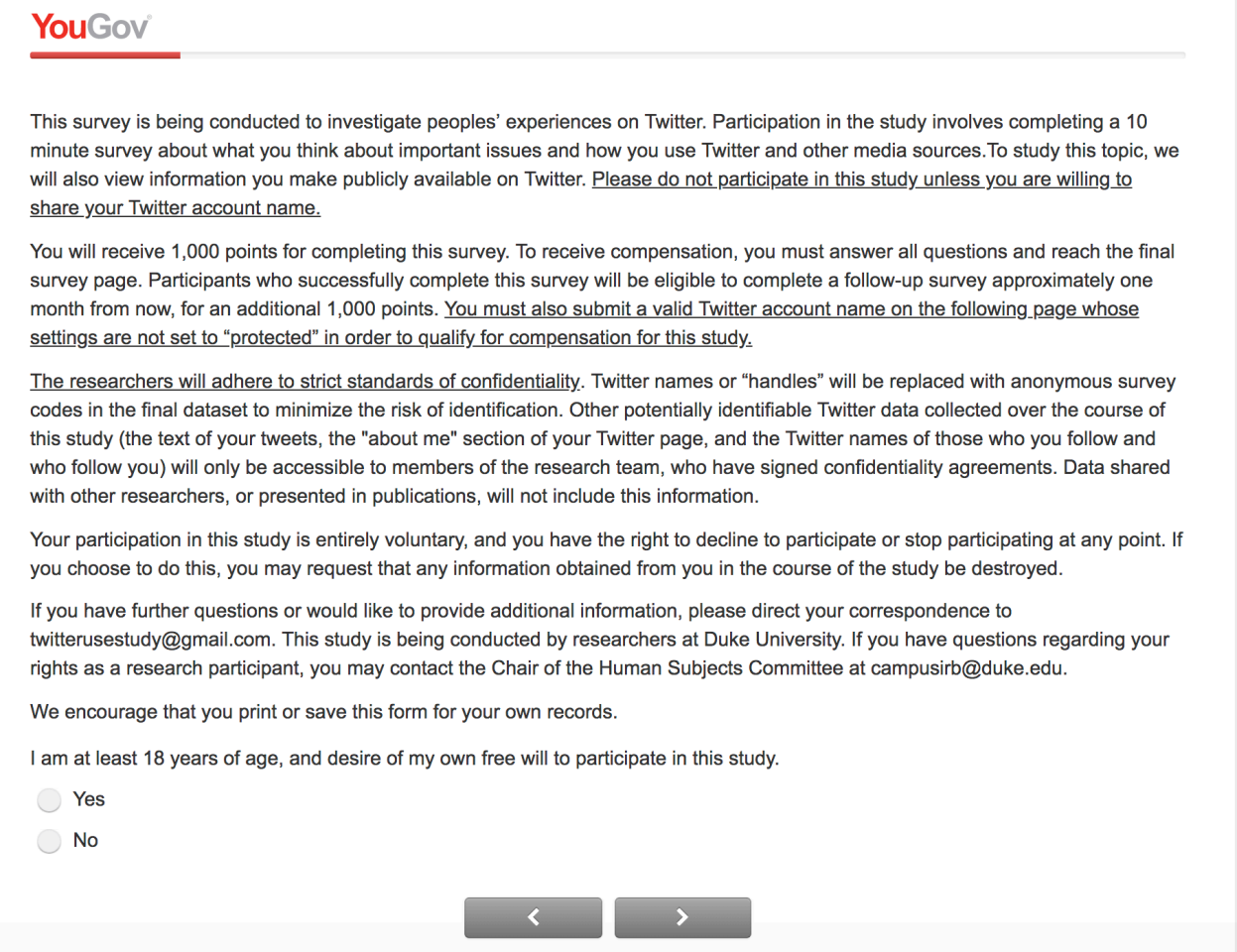

Figure S 2: Informed Consent Dialogue, Reprinted from Bail et al. (2018)

1,754 respondents agreed and completed the entire pre-treatment survey. 500 respondents began—but did not complete—the pre-treatment survey, and 285 respondents refused to complete the pre-treatment survey. Of the 1,754 respondents that completed the first survey, 102 were removed by YouGov's quality algorithm, which eliminates respondents who complete the survey within a time frame that is deemed impossible by the algorithm. This resulted in an initial sample of 1,652 respondents.

One month later, respondents who were not eliminated for one of the reasons described in the next section were invited to complete the post-treatment survey. In the informed consent for the post-treatment survey, participants were reminded "One month ago, you completed a survey to investigate your experiences on Twitter…", and were informed that they were now invited to complete a 10 minute follow up survey about "what you think about important issues and how you use Twitter and other media sources."

Respondents Eliminated Before Treatment Assignment

136 of the 1,652 respondents who completed the first survey were excluded from subsequent analyses because they did not present a valid Twitter handle or username that could be accessed via Twitter's Application Programming Interface. Forty-five respondents were excluded because they provided poor quality data, indicated by providing the same answer to ten consecutive questions that were randomized according to whether the respondent was asked to agree with a liberal or conservative-leaning statement (an additional twelve respondents were later excluded because they did the same during the final post-treatment survey).

An advantage of Bail et al.'s research design is that the authors were able to cross-validate survey responses with behavioral and demographic information available from the Twitter profiles and messages of respondents to the first survey. Forty-four respondents were excluded because they did not follow any accounts on Twitter, and therefore did not satisfy the screening criterion that participants be active Twitter users who "regularly log on to read messages from other Twitter accounts." Respondents were also eliminated if the demographic information they provided in the survey conflicted with at least two demographic variables that were observable on the respondent's Twitter page (age, gender, race, and geographic location). Some of these respondents were excluded because of the aforementioned exclusion rules, but an additional 74 respondents were dropped because they provided highly inconsistent information in the survey and in their Twitter profile. Four additional respondents were excluded because they provided an account of a famous person instead of their own Twitter account. Bail et al. operationalized fame as having more than 50,000 followers. Because it was theoretically possible that a respondent in the study could have a large number of followers, Bail et al. cross-referenced demographic information from the Twitter account in question with that reported in the survey and identified significant discrepancies which further suggested that these responses were non-valid.

Causal Interference

Another advantage of Bail et al.'s research design is that they were able to collect social network data from each respondent's Twitter account in order to mitigate the risk of causal interference within the original study's field experiment. After removing respondents who were excluded for reasons described in the previous section of this document, the authors identified 136 respondents in their sample who followed—or were followed by—at least one other respondent in the study. As Figure S3 (on the next page) shows, 90 of these people were part of network components that included at least two other participants in the study. All respondents that were part of such components were excluded. Of the remaining 46 people who were connected to only one other person in the survey population, one person in the pair was randomly dropped.

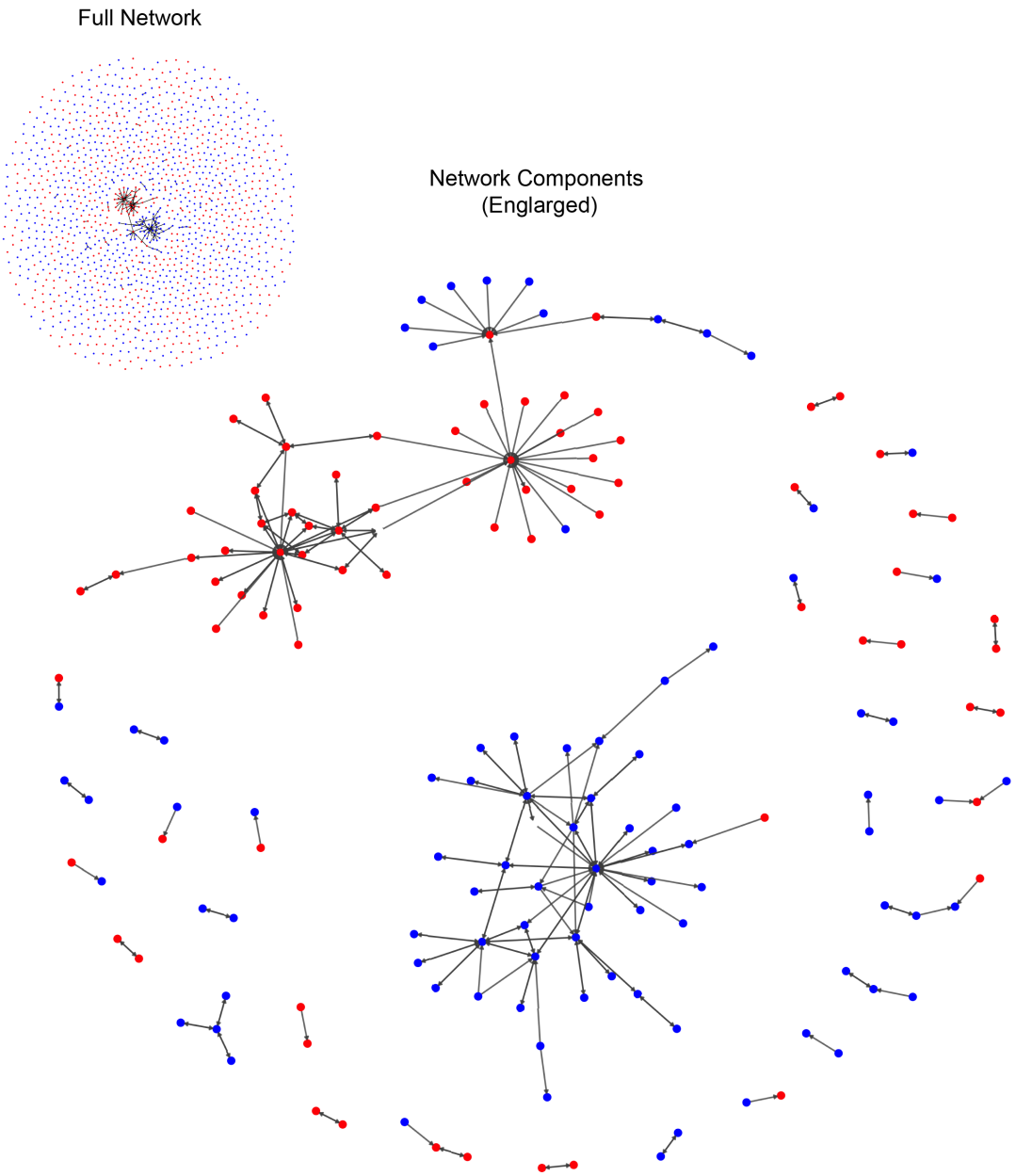

Figure S 3: Network Interference, Reprinted from Bail et al. (2018)

Descriptive Characteristics of Final Study Population

Bail et al.'s final sample included 1,239 respondents. Table S1 below compares the final sample to data from the 2016 American Community Survey, which is available here. Data on state populations for 2016 were collected from the U.S. Census, and are available at this link. Although the sample was drawn to be generalizable to partisan Twitter users, this table shows that the sample also closely approximates the adult population distribution across geographic regions of the United States, races and ethnicities, and gender. Table S1 also provides descriptive characteristics of both Democratic and Republican respondents in the study. Given the nonprobability sampling design, the sample might differ from the general population of partisan Twitter users in other unknown ways.

Table S 1: Comparison of Demographic Characteristics of Respondents in Study Sample to U.S. Census/American Community Survey, Reprinted from Bail et al. (2018)

Variable	National Mean	Study Mean	p	Study Dem. Mean	Study Rep. Mean	p
Age	37.84	50.49	0.87	50.31	50.72	0.64
Female	0.51	0.52	0.48	0.55	0.48	0.02
White	0.70	0.84	0.00	0.78	0.92	0.00
Asian	0.05	0.01	0.00	0.01	0.01	0.38
Black	0.12	0.06	0.00	0.10	0.02	0.00
Native American	0.01	0.01	0.89	0.01	0.01	0.77
Other	0.05	0.05	0.83	0.06	0.04	0.04
Mixed	0.06	0.02	0.00	0.03	0.01	0.01
New England	0.03	0.03	1.00	0.04	0.03	0.55
Mid-Atlantic	0.13	0.15	0.00	0.17	0.13	0.05
East North Central	0.14	0.14	0.87	0.15	0.13	0.62
West North Central	0.06	0.06	0.43	0.06	0.05	0.55
South Atlantic	0.20	0.22	0.05	0.20	0.24	0.15
East South Central	0.06	0.07	0.24	0.06	0.08	0.29
West South Central	0.12	0.10	0.01	0.08	0.12	0.01
Mountain	0.07	0.06	0.02	0.05	0.06	0.56
Pacific	0.18	0.17	0.40	0.19	0.15	0.11

Field Experiment

Respondents assigned to the treatment condition were invited to follow a Twitter bot, described in additional detail below, that exposed them to messages from opinion leaders from the opposing political party. This invitation was sent on October 21, 2017, roughly one week after they were recruited to complete the first survey described above. The one week buffer between the initial survey and treatment was intended to decrease the likelihood that respondents would become aware of the purpose of the experiment (see additional discussion of experiment effects below). The invitation to those in the treatment condition was as follows: "Recently you completed a survey for YouGov about how often you use Twitter. You have been randomly selected for an opportunity to receive up to 10,000 points for completing an additional task related to that survey. Participation in this portion of the study will involve following a Twitter account created as a part of the study that will tweet 24 messages per day for one month." The invitation then stated that respondents must follow the bot for an entire month in order to receive compensation and provided details about how respondents could receive additional incentives for correctly answering questions about the bot's tweets (as we describe in further detail below). Finally, the invitation clarified that participation in this portion of the study was entirely voluntary, and told respondents that they had the right to decline to participate or stop participating at any point. The final paragraph of the invitation invited them to contact the Human Subjects Committee at the first author's institution if they have any questions regarding their rights as a research participant.

If respondents accepted the invitation they were redirected to another web page that included a link to follow either the study's liberal or conservative bot, depending upon their self-reported party identification. This page informed them they would earn the equivalent of $11 for following the bot and up to an additional $18 for successfully answering questions about the content of the messages retweeted by the bot during surveys that would follow each week. Each bot was given a non-descript name that did not prime the political ideology of opinion leaders that it retweeted. We are unable to report the names here because data collection continues for follow-up research and disclosure of the twitter handles could be used to identify respondents in the study who engaged with the bot by commenting on, liking, or retweeting its retweets. For the first few days in which respondents began following the bots, only pictures of nature landscapes were retweeted in order to further mask the purpose of the study.

Design of Twitter Bots

The two Twitter bots created for Bail et al.'s study were designed as follows. First, the authors built upon Barbera et al.'s (2015) ideological scoring method for Twitter users. They began by collecting the Twitter IDs, or "handles," for all presidential candidates and members of the House and Senate as of August, 5 2017. Bail et al. then scraped the names of all people who these elected officials follow from Twitter's Application Programming Interface, which yielded a total sample of 636,737 Twitter accounts. Next, the researchers eliminated all those who were not followed by at least 15 of the aforementioned elected officials. They then conducted a correspondence analysis on the resultant adjacency matrix, and used the first principal component to create a liberal/conservative score for all of those in this "opinion leader" network. Bail et al. binned this scale into seven quantiles, and dropped those in the fourth, centrist, quantile. The liberal bot randomly retweeted messages from opinion leaders in the first, second, and third quantile produced during the preceeding 24 hours, and the conservative bot randomly retweeted messages from opinion leaders in the fifth, six, and seventh quantiles during the preceeding 24 hours.

Bail et al. took several additional steps to improve the ideological scores used to create the bots. First, they eliminated all U.S. government agencies, since most of these retweeted non-partisan messages that would dilute treatment. Second, they eliminated all accounts that were administered by for-profit U.S. corporations, though they did not eliminate non-profit organizations, think tanks, or other nonprofit groups. Third, Bail et al. eliminated a small number of accounts that were controlled by elected officials outside the United States.

Despite these steps, pilot analyses of the ideological continuum consistently identified a small number of elected officials who were misclassified according to the ideological scoring measure described above. Each of these individuals were very high profile opinion leaders such as Mitch McConnell and John McCain, who had very large followings that include a large number of non-Republicans, which made them centrists instead of conservatives in the original analysis. Bail et al. thus reclassified the small number of elected officials who were mistakenly identified by assigning them a random ideological score between the first and second quantile of opinion leaders that defined their party using the first principal component measure described above.

The liberal and conservative Twitter bots created by the research team were both hosted on an Amazon EC2 server. Every hour, the program randomly drew a message produced by an elected official or opinion leader from the previous 24 hours from one of the two samples. During three of the four weeks of the study period, the bot retweeted a different animal picture at two random times each day. Respondents were asked if they could identify these animals to monitor treatment compliance alongside substantive questions about the bot's tweets. For a further description of compliance measures employed, see Bail et al. (2018)

Ethics and Protection of Human Subjects

Bail et al.'s study was approved by the Institutional Review Boards at Duke University and New York University. All respondents digitally signed an informed consent dialogue before they participated in the research. Because open-ended questions in the authors' pilot study indicated Republican respondents might have anti-intellectual sentiment that could create measurement error, the informed consent dialogue did not state that the research was being conducted by academic researchers, though the name of the first author's university was listed in the penultimate paragraph alongside instructions about how to contact the Institutional Review Board with complaints about the research (see informed consent screenshot above for exact language used). No such complaints were received.

Though most Twitter data is publicly available for academic research, Bail et al.'s study linked such data to confidential survey data. Because such data are highly sensitive, we will not publicly release the names, twitter handles, or numeric ids of any respondents in the current study of IRA influence. Nor do we make the content of their tweets, the names of the people they follow, or the names of the people who follow them publicly available. Instead, the public release of data for this study of IRA influence will include a variable that describes each respondent's a) number of twitter followers, b) the number of people they follow, and c) a measure of the ideological heterogeneity of their Twitter network which is described in additional detail below. We coarsen these variables into increments of 50 within the public release data in order to prevent them being used to identify respondents in our study.

Outcome Variables

Liberal Conservatism Scale

Bail et al.'s study employed a variation of the "ideological consistency scale" developed by previous studies (Dimock and Carroll 2014). The scale, which asks respondents to agree or disagree with a series of twenty statements about social policies worded to favor either liberal or conservative views, was previously included in sixteen nationally representative surveys. Bail et al. made two important modifications to this scale. Instead of a binary choice between liberal and conservative options for each policy statement, they used a seven-point response scale, since allowing respondents to indicate strength or extremity of opinion provides a more accurate measure of ideological polarization (Fiorina, Abrams, and Pope 2006; Hill and Tausanovitch 2015). Second, instead of asking respondents to read twenty questions, Bail et al. randomly selected five liberal versions of each policy statement and five conservative versions.

Thus, Bail et al.'s survey asked respondents to agree or disagree with the following statements on a seven point scale from "strongly disagree" to "strongly agree."

1) "Stricter environmental laws and regulations cost too many jobs and hurt the economy."
2) "Government regulation of business is necessary to protect the public interest."
3) "Poor people today have it easy because they can get government benefits without doing anything in return."
4) "Immigrants today strengthen our country because of their hard work and talents."
5) "Government is almost always wasteful and inefficient."
6) "The best way to ensure peace is through military strength."
7) "Racial discrimination is the main reason why many black people can't get ahead these days."
8) "The government today can't afford to do much more to help the needy."
9) "Business corporations make too much profit."
10) "Homosexuality should be accepted by society."

As previously mentioned, half of these statements are worded in a manner that is designed to appeal to liberals (#2,#4,#7,#9,#10), and the other half are intended to appeal to conservatives (#1,#3,#5,#6,#8). Question order was randomized in both the pre and post-treatment surveys.

The code below was used to create the outcome measure for the current study of IRA influence. Liberal questions were reverse-coded such that negative values on the outcome indicate respondents becoming more liberal and positive values indicate respondents becoming more conservative. We calculate the mean score on this ten-item index for the first and second survey wave, and our models predict change between the two survey waves.

```
load("For Troll Influence Models.Rdata")

#invert questions that prime liberal values
 troll_analysis_data$government_should_regulate_businesses_wave_1<-
   8-troll_analysis_data$government_should_regulate_businesses_wave_1
 troll_analysis_data$racial_discrimination_hurts_black_people_wave_1<-
   8-troll_analysis_data$racial_discrimination_hurts_black_people_wave_1
 troll_analysis_data$immigrants_strengthen_country_wave_1<-
   8-troll_analysis_data$immigrants_strengthen_country_wave_1
 troll_analysis_data$corporations_make_too_much_profit_wave_1<-
   8-troll_analysis_data$corporations_make_too_much_profit_wave_1
 troll_analysis_data$homosexuality_should_be_accepted_wave_1<-
```

```
  8-troll_analysis_data$homosexuality_should_be_accepted_wave_1
troll_analysis_data$government_should_regulate_businesses_wave_5<-
  8-troll_analysis_data$government_should_regulate_businesses_wave_5
troll_analysis_data$racial_discrimination_hurts_black_people_wave_5<-
  8-troll_analysis_data$racial_discrimination_hurts_black_people_wave_5
troll_analysis_data$immigrants_strengthen_country_wave_5<-
  8-troll_analysis_data$immigrants_strengthen_country_wave_5
troll_analysis_data$corporations_make_too_much_profit_wave_5<-
  8-troll_analysis_data$corporations_make_too_much_profit_wave_5
troll_analysis_data$homosexuality_should_be_accepted_wave_5<-
  8-troll_analysis_data$homosexuality_should_be_accepted_wave_5


# #create average score by wave
# # mean issue scale (wave 1)
troll_analysis_data$substantive_ideology_scale_wave_1<-rowMeans(troll_analysis_data[,c(
  "government_should_regulate_businesses_wave_1",
  "racial_discrimination_hurts_black_people_wave_1",
  "immigrants_strengthen_country_wave_1",
  "corporations_make_too_much_profit_wave_1",
  "homosexuality_should_be_accepted_wave_1",
  "government_wasteful_inefficient_wave_1",
  "poor_people_have_it_easy_wave_1",
  "government_cannot_afford_to_help_needy_wave_1",
  "best_way_peace_military_strength_wave_1",
  "stricter_environmental_laws_damaging_wave_1")], na.rm=TRUE)

# # mean issue scale (wave 5)
troll_analysis_data$substantive_ideology_scale_wave_5<-rowMeans(troll_analysis_data[,c(
  "government_should_regulate_businesses_wave_5",
  "racial_discrimination_hurts_black_people_wave_5",
  "immigrants_strengthen_country_wave_5",
  "corporations_make_too_much_profit_wave_5",
  "homosexuality_should_be_accepted_wave_5",
  "government_wasteful_inefficient_wave_5",
  "poor_people_have_it_easy_wave_5",
  "government_cannot_afford_to_help_needy_wave_5",
  "best_way_peace_military_strength_wave_5",
  "stricter_environmental_laws_damaging_wave_5")], na.rm=TRUE)

with(troll_analysis_data, cor(substantive_ideology_scale_wave_1,
                              substantive_ideology_scale_wave_5,
                              use = "complete.obs"))

troll_analysis_data$substantive_ideology_change<-0
#loop to code affect change by party id of respondent
for(i in 1:nrow(troll_analysis_data)){
  #democrats
  if(troll_analysis_data$party_id_wave_1[i]==1){
    #invert scale for liberals
    troll_analysis_data$substantive_ideology_change[i]<-
      -(troll_analysis_data$substantive_ideology_scale_wave_5[i]-
      troll_analysis_data$substantive_ideology_scale_wave_1[i])
```

```
   }
   if(troll_analysis_data$party_id_wave_1[i]==2){
#     #do not invert for republicans
     troll_analysis_data$substantive_ideology_change[i]<-
       troll_analysis_data$substantive_ideology_scale_wave_5[i]-
       troll_analysis_data$substantive_ideology_scale_wave_1[i]
   }
 }

#save(troll_analysis_data, file = "For Troll Influence Models.Rdata")
```

Social Distance

The code below was used to create the measure of partisan social distance. Respondents were asked the extent to which they would agree or disagree with the following statements about the opposite political party: "I would be unhappy if someone in my immediate family married a (Democrat/Republican)", "I would be unhappy if I had to spend time socializing with a (Democrat/Republican)", and "I would be unhappy if I had to work closely with a (Democrat/Republican)". Responses were measured on a 7-point scale, ranging from Strongly Disagree to Strongly Agree. To create a measure of social distance we calculate the difference between each respondent's average among these three questions in the first and second waves of the study.

```
 load("For Troll Influence Models.Rdata")

#create wave 1 social distance mean index
 troll_analysis_data$social_distance_index_wave_1 <-
   rowMeans(troll_analysis_data[,c("unhappy_family_married_other_party_wave_1",
                                   "unhappy_socializing_other_party_wave_1",
                                   "unhappy_work_close_other_party_wave_1")],
            na.rm=T)

 troll_analysis_data$social_distance_index_wave_5 <-
   rowMeans(troll_analysis_data[,c("unhappy_family_married_other_party_wave_5",
                                   "unhappy_socializing_other_party_wave_5",
                                   "unhappy_work_close_other_party_wave_5")],
            na.rm=T)


 troll_analysis_data$social_distance_change <-
   troll_analysis_data$social_distance_index_wave_5-
   troll_analysis_data$social_distance_index_wave_1
#
# save(troll_analysis_data, file = "For Troll Influence Models.Rdata")
```

Feeling Thermometers

The code below was used to create the measure of change in out-group feeling thermometer ratings. Respondents were asked to rate the Democratic and Republican parties on a scale from 0-100, where higher values indicate more positive, or warmer, feelings. The change in feeling thermometer rates was calculated as a difference in differences: that is, the difference between the change in a respondent's rating of the opposing party across survey waves and the change in their rating of their own party. This approach enables us to account for individual-level differences in how respondents use the 101-point thermometer scale, in addition to providing a relative, rather than absolute, measure of attitudes toward the opposing political party.

```
 load("For Troll Influence Models.Rdata")
#
 # Out-Party Feeling Thermomter Change
 troll_analysis_data$out_party_ft_change <-
   with(troll_analysis_data,
        ifelse(party_id_wave_1 == 1,
               thermometer_republicans_wave_5 -
                 thermometer_republicans_wave_1,
               ifelse(party_id_wave_1 == 2,  thermometer_democrats_wave_5 -
                        thermometer_democrats_wave_1, NA)))

# # In-Party Feeling Thermomter Change
 troll_analysis_data$in_party_ft_change <-
   with(troll_analysis_data,
        ifelse(party_id_wave_1 == 2,
               thermometer_republicans_wave_5 -
                 thermometer_republicans_wave_1,
               ifelse(party_id_wave_1 == 1,  thermometer_democrats_wave_5 -
                        thermometer_democrats_wave_1, NA)))

# Out-Party(Week5 - Week1) - In-Party(Week 5 - Week 1)
 troll_analysis_data$ft_change <- with(troll_analysis_data,
                     out_party_ft_change - in_party_ft_change)

# save(troll_analysis_data, file="For Troll Influence Models.Rdata")
```

Missing Data

We employed imputation to address a small amount of missing data in Bail et al.'s pre-treatment survey—particularly the variables that describe respondents' income (7% missing). We do not impute any missing data for outcome variables.

```
load("For Troll Influence Models.Rdata")

#Main model: treated = any interaction, indirect OR direct

#Drop respondents who interacted with trolls before the study
troll_analysis_data <- troll_analysis_data %>% filter(ind_intbefore != 1)

#create indicator of Troll Exposure
troll_analysis_data$treated <- troll_analysis_data$ind_intduring

#Impute missing data

#Drop wave 2 incompletes
troll_analysis_data <- troll_analysis_data[!is.na(troll_analysis_data$ft_change) |
                                is.na(troll_analysis_data$social_distance_change) |
                                is.na(troll_analysis_data$substantive_ideology_change) |
                                is.na(troll_analysis_data$ideology_change) |
                                is.na(troll_analysis_data$change_num_oplead_follows) |
                                is.na(troll_analysis_data$echo_chamber_change),]

#Subset only variables used in final BCF models below
```

```
  for_imputation<-troll_analysis_data[,c(
          "treated",
          "republican_wave_1",
          "how_often_visit_twitter_wave_1",
          "newsint_wave_1",
          "birth_year",
          "male",
          "college",
          "white",
          "family_income",
          "northeast",
          "south",
          "west",
          "treat",
          "ft_change",
          "social_distance_change",
          "substantive_ideology_change",
          "ideology_change",
          "change_num_oplead_follows",
          "echo_chamber_change"
        )]

#Impute
  library(mice)
  imputed_datasets <- mice(for_imputation,m=15,seed=352)
  imputed_data <- mice::complete(imputed_datasets, action=15)

save(imputed_data, file="Final Imputed Data.Rdata")
```

Frequency of IRA Account Interaction

Our main analysis considers the effect of respondents' interactions with trolls, including both direct engagements and indirect engagements. We operationalize *direct engagement* as mentioning a troll in a tweet (including both retweets of trolls and direct replies to trolls), liking a troll's tweet, or liking a tweet that mentions a troll. Overall, 11.3% of our respondents had some type of direct engagement with a troll on Twitter throughout the entire range of data we were able to attain from Twitter. Within the same period, mentioning a troll was the most common type of direct engagement, with 6.7% of respondents engaging in this activity, closely followed by 6.4% who ever liked a troll tweet. It was less common to directly engage with troll content by liking a tweet that mentioned a troll (<1% of respondents).

We operationalize interactions as including both these direct engagements as well as *indirect engagements* that were likely to lead to respondents viewing troll tweets: 1) following trolls, and 2) following a person who mentions a troll. Each of these actions make it more likely that troll content appeared on our respondents' timelines. Unfortunately, Twitter did not respond to our request to provide data indicating exactly who viewed the troll's tweets. Yet the relatively large amount of respondents who had friends mention a troll (10.73%), indicates this indirect exposure helps to include a wider understanding of being exposed to troll content on Twitter. Following a troll was much less common (1.05% of respondents). Overall, 235 respondents (almost 19% of our sample) had some type of interaction (direct or indirect engagement) with a troll on Twitter throughout the entire range of data we were able to attain from Twitter's Application Programming Interface.

Table S2 illustrates the frequency of troll interactions before and during the main survey waves analyzed in the central models presented in the main text of our manuscript. During the one-month period of the study, a total of 17 respondents (1.37%) directly engaged with trolls for the first time, and 46 respondents (3.71) otherwise interacted with trolls for the first time. The pattern of troll interactions followed the trend

seen in the overall time period, with the majority of direct engagements stemming from liking a troll tweet and the majority of interactions stemming from having a friend mention a troll.

Table S 2: Troll Interaction Summary Statistics

	Entire Time Period		Only During Study	
	Number	Percent	Number	Percent
Any Interaction	**235**	**18.97**	**46**	**3.71**
Any Direct Engagement	**140**	**11.30**	**17**	**1.37**
Ever Mention Troll	83	6.70	5	0.40
Ever Like Troll	79	6.38	14	1.13
Ever Like Troll Mention	10	0.81	0	0.00
Ever Follow Troll	13	1.05	6	0.48
Ever Friend Mention	133	10.73	28	2.26

Lastly, for individuals who did interact with trolls, we calculated the percent of troll interactions out of all twitter interactions, to estimate the level of treatment.

```
#For calculating the percentage of interactions that are troll interactions

library(tidyverse)

#get all tweet ids from these three datasets:
#favorites (direct and favorites of mentions), mentions, and retweets
load("Russian Troll Favorites.Rdata")
load("Russian Troll Mention Favorites.Rdata")
load("Russian Troll Tweets.Rdata")

faveids <- russian_troll_favorites %>%
  select(id) %>%
  unique()

menfaveids <- russian_troll_mention_favorites %>%
  select(id) %>%
  unique()

menrtids <- russian_troll_tweets %>%
  select(id) %>%
  unique()

alltrollids <- faveids %>%
  full_join(menfaveids) %>%
  full_join(menrtids) %>%
  unique() %>%
  mutate(isTroll = TRUE)

load("Respondent Favorites.Rdata")
load("Respondent Twitter Timelines.Rdata")
head(respondent_favorites)
names(respondent_favorites)
names(respondent_tweets)

respondent_all <- respondent_favorites %>%
  mutate(respFave = 1) %>%
```

```
  full_join(respondent_tweets) %>%
  left_join(alltrollids, by="id") %>%
  mutate(isTrollNum = ifelse(is.na(isTroll), 0,
                             ifelse(isTroll, 1,
                                    -9)))

byuser <- respondent_all %>%
  group_by(id_str) %>%
  summarize(nTroll = sum(isTrollNum),
            nEngage = n(),
            fracTroll = sum(isTrollNum)/n())

usersWithTrolls = byuser %>%
  filter(nTroll > 0)

print(mean(usersWithTrolls$fracTroll))
```

In all, on average 0.098248% of our treated respondents' engagements involve trolls.

Predicting IRA Account Interaction

The code chunk below produces Figure S1 from the main text of the manuscript. We impute missing data for the independent variables in the model.

```
load("For Troll Influence Models.Rdata")

#create indicator of Troll Exposure - includes ALL interaction, direct and indirect
troll_analysis_data <- troll_analysis_data %>%
                       mutate(binary_engagement = ifelse(ind_intbefore > 0, 1, 0))

#Standardize continuous variables

troll_analysis_data$birth_year<-
  (troll_analysis_data$birth_year -
     mean(troll_analysis_data$birth_year))/
  sd(troll_analysis_data$birth_year)

troll_analysis_data$friends_count_wave_1<-
  (troll_analysis_data$friends_count_wave_1-
 mean(troll_analysis_data$friends_count_wave_1))/
  sd(troll_analysis_data$friends_count_wave_1)

troll_analysis_data$family_income<-
  (troll_analysis_data$family_income-
  mean(troll_analysis_data$family_income,na.rm=TRUE))/
  sd(troll_analysis_data$family_income, na.rm=TRUE)

troll_analysis_data$how_often_visit_twitter_wave_1<-
  (troll_analysis_data$how_often_visit_twitter_wave_1-
  mean(troll_analysis_data$how_often_visit_twitter_wave_1))/
  sd(troll_analysis_data$how_often_visit_twitter_wave_1)


# Impute Missing Data
```

```r
#Subset only variables used in final BCF models below
  for_imputation_count <-troll_analysis_data[,c(
     "binary_engagement",
     "republican_wave_1",
     "birth_year",
     "male",
     "college",
     "family_income",
     "white",
     "northeast",
     "north_central",
     "south",
     "newsint_wave_1",
     "percent_co_party",
     "friends_count_wave_1",
     "how_often_visit_twitter_wave_1")]


#Impute
  library(mice)
  imputed_datasets_count <- mice(for_imputation_count,m=15,seed=352)
  imputed_data_count <- mice::complete(imputed_datasets_count, action=15)







#Binomial regression Model
 engagement_model<-glm(binary_engagement ~
                                republican_wave_1+
                                birth_year+
                                male+
                                college +
                                family_income +
                                white +
                                northeast +
                                north_central +
                                south +
                                newsint_wave_1+
                                percent_co_party+
                                friends_count_wave_1+
                                how_often_visit_twitter_wave_1,
                                family = binomial,
                                data = imputed_data_count)


#Make Coefficient Plot

engagement_plot<-as.data.frame(summary(engagement_model)[[12]])
engagement_plot$variable<-row.names(engagement_plot)
engagement_plot$variable<-factor(engagement_plot$variable,
                                levels=c(
```

```
                                "north_central",
                                "college",
                                "male",
                                "birth_year",
                                "northeast",
                                "south",
                                "family_income",
                                "friends_count_wave_1",
                                "white",
                                "republican_wave_1",
                                "how_often_visit_twitter_wave_1",
                                 "newsint_wave_1",
                                "percent_co_party"
                                 ),
                                 labels=c(
                                "North Central",
                               "College Degree",
                                "Male",
                                "Year of Birth",
                                "Northeast",
                                "South",
                                "Family Income",
                                "# Twitter Accounts Followed",
                                "White",
                                "Republican",
                                "Frequency of Twitter Use",
                                "Interest in Politics",
                                 "% Co-Partisans Followed"
                                 ))

#Drop intercept
engagement_plot<-engagement_plot[2:nrow(engagement_plot),]
names(engagement_plot)<-c("estimate", "se","z","p","variable")

#Plot
library(ggplot2)

interval1 <- -qnorm((1-0.9)/2)  # 90% multiplier
interval2 <- -qnorm((1-0.95)/2)  # 95% multiplier

p<-ggplot(engagement_plot)+
  geom_hline(yintercept = 0, colour = gray(1/2), lty = 2)+
  geom_linerange(aes(x = reorder(variable, estimate), ymin = estimate - se*interval1,
                     ymax = estimate + se*interval1),
                 lwd = 1, position = position_dodge(width = 1/2),
                 colour="skyblue")+
  geom_linerange(aes(x = variable, y = estimate, ymin = estimate - se*interval2,
                     ymax = estimate + se*interval2),
                 lwd = .5, position = position_dodge(width = 1/2),
                 colour="skyblue")+
    geom_point(aes(x=variable, y=estimate),
              position = position_dodge(width = 1/2),
              size=2, colour="#71084D")+
```

```
  theme(axis.text=element_text(size=10, face="bold",colour="black"),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        panel.background=element_blank(),
        axis.title=element_text(size=10, colour="black"),
        legend.position="none",
        legend.key = element_blank(),
        legend.title=element_blank())+
  labs(x="",y="")+
  coord_flip()
p
 ggsave(p, file="Binomial_Regression_Plot.eps",
        width=10, height=6, device = "eps", dpi = 300)
```

Bayesian Causal Forest Models

The code chunk below creates Figure 2 in the main text of the manuscript as well as the figures that describe causal heterogeneity by covariate (Figures 3-5) and additional figures that appear at the end of the Supplementary Materials.

```
load("Final Imputed Data.Rdata")

#Make outcomes have mean 0 and standard deviation 1 for comparable scales in plot
imputed_data$ft_change<-
  (imputed_data$ft_change-
     mean(imputed_data$ft_change))/
  sd(imputed_data$ft_change)

imputed_data$social_distance_change<-
  (imputed_data$social_distance_change-
     mean(imputed_data$social_distance_change))/
  sd(imputed_data$social_distance_change)

imputed_data$substantive_ideology_change<-
  (imputed_data$substantive_ideology_change-
     mean(imputed_data$substantive_ideology_change))/
  sd(imputed_data$substantive_ideology_change)

imputed_data$ideology_change<-
  (imputed_data$ideology_change-
     mean(imputed_data$ideology_change))/
  sd(imputed_data$ideology_change)

imputed_data$change_num_oplead_follows<-
  (imputed_data$change_num_oplead_follows-
     mean(imputed_data$change_num_oplead_follows))/
  sd(imputed_data$change_num_oplead_follows)

imputed_data$echo_chamber_change<-
  (imputed_data$echo_chamber_change-
     mean(imputed_data$echo_chamber_change))/
  sd(imputed_data$echo_chamber_change)
```

```r
#binarize variables for easier interpretation

imputed_data$over_65<-
  ifelse(imputed_data$birth_year<1952, 1,0)
imputed_data$frequent_twitter_user<-
  ifelse(imputed_data$how_often_visit_twitter_wave_1==3, 1, 0)
imputed_data$high_news_interest<-
  ifelse(imputed_data$newsint_wave_1==4, 1, 0)
imputed_data$high_income<-ifelse(imputed_data$family_income>9, 1, 0)

#propensity score matching

library(MatchIt)
m.out <- matchit(treated ~
                 republican_wave_1+
                 frequent_twitter_user+
                 high_news_interest+
                 over_65+
                 male+
                 college+
                 white+
                 high_income+
                 northeast+
                 south+
                 west+
                 treat,
                 data = imputed_data,
                 method = "nearest")

#Function to run BCF Models and Plot Heterogeneous Treatment Effects

library(bcf)
library(tidyr)
library(ggplot2)


bcf_machine<-function(y, name){

  bcf_fit<-bcf(y,
               imputed_data$treated,
               x_control=as.matrix(imputed_data[,c(
                 "republican_wave_1",
                 "frequent_twitter_user",
                 "high_news_interest",
                 "over_65",
                 "male",
                 "white",
                 "college",
                 "high_income",
                 "northeast",
                 "south",
                 "west",
                 "treat")
```

```
                ]),
              x_moderate = as.matrix(imputed_data[,c(
                "republican_wave_1",
                "frequent_twitter_user",
                "high_news_interest",
                "over_65",
                "male",
                "white",
                "college",
                "high_income",
                "northeast",
                "south",
                "west",
                "treat")
                ]),
              pihat=m.out$distance,
              nburn=5000,
              nsim=20000,
              update_interval=10000)

#extract individual estimates
tau_post<-bcf_fit$tau
tauhat<-colMeans(tau_post)

#make plot dataframe for heterogeneous treatment effects
imputed_data$individual_effects<-tauhat
forggplot<-imputed_data[,c(1,2,6,7,8,20:24)]

plotter<-forggplot %>%
gather(Variable, Value,  -individual_effects)

#add to plot the individual estimates by the treatment indicator

  plotter$Variable<-factor(plotter$Variable,
                           levels=c(
                             "treated",
                             "republican_wave_1",
                             "high_news_interest",
                             "frequent_twitter_user",
                             "over_65",
                             "male",
                             "college",
                             "high_income",
                             "white"
                           ),
                           labels=c(
                             "IRA Account Interaction (1=Yes, 0=No)",
                             "Republican (1=Yes, 0=No)",
                             "Interest in Politics (1=High, 0=Not High)",
                             "Frequency of Twitter Use (1=High, 0=Not High)",
                             "Over Age 65 (1=Yes, 0=No)",
                             "Male (1=Yes, 0=No)",
                             "College Degree (1=Yes,0=No)",
```

```
                            "High Income (1=Yes, 0=No)",
                            "White (1=Yes, 0=No)"))

p1 <- ggplot(subset(plotter, Variable == "Republican (1=Yes, 0=No)"),
             aes(y = individual_effects, x = as.factor(Value))) +
  geom_jitter(size=.5)+
  geom_boxplot(outlier.shape = NA, fill = "#d53e4f")+
  scale_x_discrete(labels=c("0" = "Democrat", "1" = "Republican")) +
  ggtitle("Political Party") +
  theme_minimal()+
  theme(axis.title.x=element_blank(),
        axis.title.y=element_blank(),
        axis.text.y = element_blank(),
        legend.position="none",
        plot.title = element_text(hjust = 0.5, face="bold"))

p2 <- ggplot(subset(plotter, Variable == "Interest in Politics (1=High, 0=Not High)"),
             aes(y = individual_effects, x = as.factor(Value))) +
  geom_jitter(size=.5)+
  geom_boxplot(outlier.shape = NA, fill = "#f46d43") +
  ggtitle("Interest in Politics")+
  scale_x_discrete(labels=c("0" = "Not High", "1" = "High")) +
  theme_minimal()+
  theme(axis.title.y=element_blank(),
        axis.title.x=element_blank(),
        axis.text.y = element_blank(),
        legend.position="none",
        plot.title = element_text(hjust = 0.5, face="bold"))

p3 <- ggplot(subset(plotter,
                    Variable == "Frequency of Twitter Use (1=High, 0=Not High)"),
             aes(y = individual_effects, x = as.factor(Value))) +
  geom_jitter(size=.5)+
  geom_boxplot(outlier.shape = NA, fill = "#ffffbf") +
  theme_minimal()+
  ggtitle("Frequency of Twitter Use")+
  scale_x_discrete(labels=c("0" = "Not High","1" = "High")) +
  theme(axis.title.y=element_blank(),
        axis.title.x=element_blank(),
        legend.position="none",
        plot.title = element_text(hjust = 0.5, face="bold"))

p4 <- ggplot(subset(plotter, Variable == "Over Age 65 (1=Yes, 0=No)"),
             aes(y = individual_effects, x = Value, group = Value))+
  geom_jitter(size=.5)+
  geom_boxplot(outlier.shape = NA, fill = "#5aae61") +
  theme_minimal()+
  scale_x_discrete(limits = c("1"="Over 65", "0"="Under 65")) +
  ggtitle("Year of Birth") +
  theme(axis.title.y=element_blank(),
        axis.title.x=element_blank(),
        axis.text.y = element_blank(),
        legend.position="none",
```

```
        plot.title = element_text(hjust = 0.5, face="bold"))

p5 <- ggplot(subset(plotter, Variable == "Male (1=Yes, 0=No)"),
             aes(y = individual_effects, x = as.factor(Value))) +
  geom_jitter(size=.5)+
  geom_boxplot(outlier.shape = NA, fill = "#4393c3")+
  scale_x_discrete(labels=c("0" = "Female", "1" = "Male")) +
  theme_minimal()+
  ggtitle("Sex") +
  theme(axis.title.x=element_blank(),
        axis.title.y=element_blank(),
        axis.text.y = element_blank(),
        legend.position="none",
        plot.title = element_text(hjust = 0.5, face="bold"))

p6 <- ggplot(subset(plotter, Variable == "College Degree (1=Yes,0=No)"),
             aes(y = individual_effects, x = as.factor(Value))) +
  geom_jitter(size=.5)+
  geom_boxplot(outlier.shape = NA, fill = "#313695")+
  scale_x_discrete(labels=c("0" = "No College Degree",
                            "1" = "College Degree")) +
  theme_minimal()+
  ggtitle("Education") +
  theme(axis.title.x=element_blank(),
        axis.title.y=element_blank(),
        legend.position="none",
        plot.title = element_text(hjust = 0.5, face="bold"))

p7 <- ggplot(subset(plotter, Variable == "High Income (1=Yes, 0=No)"),
             aes(y = individual_effects, x = as.factor(Value))) +
  geom_jitter(size=.5)+
  geom_boxplot(outlier.shape = NA, fill = "#762a83") +
  theme_minimal()+
  ggtitle("Family Income")+
  scale_x_discrete(labels=c("1" = "High Income","0"= "Low Income")) +
  theme(axis.title.y=element_blank(),
        axis.title.x=element_blank(),
        axis.text.y = element_blank(),
        legend.position="none",
        plot.title = element_text(hjust = 0.5, face="bold"))

p8 <- ggplot(subset(plotter, Variable == "White (1=Yes, 0=No)"),
             aes(y = individual_effects, x = as.factor(Value))) +
  geom_jitter(size=.5)+
  geom_boxplot(outlier.shape = NA, fill ="#de77ae")+
  scale_x_discrete(labels=c("0" = "Non-White", "1" = "White")) +
  theme_minimal()+
  ggtitle("Race") +
  theme(axis.title.x=element_blank(),
        axis.title.y=element_blank(),
        axis.text.y = element_blank(),
        legend.position="none",
        plot.title = element_text(hjust = 0.5, face="bold"))
```

```
        #join all of the plots together
        library(cowplot)
        plotmaker<-plot_grid(p1, p2, p3, p4, p5, p6, p7, p8, align = "h")

        #save
        save_plot(paste0("Plot of ",name, ".png"), plotmaker,
                  ncol = 3,
                  nrow = 3,
                  base_aspect_ratio = 8/6
                  )

    #Now produce the ATT, ate, and Cate
    tau.bcf <- bcf_fit$tau
    tau.bcf.treated <- tau.bcf[,imputed_data$treated==1]

    #ATT
    #These are the posterior samples of ATT (note this is an average over UNITS)
    att_s <- rowMeans(tau.bcf.treated)
    #This is an estimate of the average treatment effect on the treated
    ATT <- mean(att_s)
    #This is the lower bound of the credible interval for the average
    # treatment effect on the treated
    ATT.lb <- as.numeric(quantile(att_s,.025))
    #this is the upper bound of this credible interval
    ATT.ub <- as.numeric(quantile(att_s,.975))
    ATT.se<-sd(att_s)

    #ate
    ate_s <- rowMeans(tau.bcf)
    ate <- mean(att_s)
    ate.lb <- as.numeric(quantile(att_s,.025))
    ate.ub <- as.numeric(quantile(att_s,.975))
    ate.se<-sd(att_s)

    #Cate
    #These are the posterior estimates of Cate (note that this
    #is an average over SAMPLES)
    cate.est <- colMeans(tau.bcf)
    cate.lb <- apply(tau.bcf,2,quantile,.025)
    cate.ub <- apply(tau.bcf,2,quantile,.975)

    effects<-data.frame(cbind(ATT, ATT.se, ATT.lb, ATT.ub, ate,
                              ate.se, ate.lb, ate.ub))

    save(effects, file=paste0("Treatment Effects for ", name, ".Rdata"))

return(bcf_fit)
}

#run function for each outcome

thermometer<-bcf_machine(imputed_data$ft_change,
                         name="Thermometer Ratings of Outgroup")
```

```
sociald<-bcf_machine(imputed_data$social_distance_change,
                     name="Desired Social Distance from Outgroup")

subideo<-bcf_machine(imputed_data$substantive_ideology_change,
                     name="Substantive Ideology Change")

ideo<-bcf_machine(imputed_data$ideology_change,
                  name="Ideology Change")

num_oplead_follows<-bcf_machine(imputed_data$change_num_oplead_follows,
                                name="Change No Opinion Leaders Followed")

echo_change<-bcf_machine(imputed_data$echo_chamber_change,
                         name="Change in Strength of Echo Chamber")


#now load results and create plot of att

load("Treatment Effects for Thermometer Ratings of Outgroup.Rdata")

thermometer_att<-effects

load("Treatment Effects for Desired Social Distance from Outgroup.Rdata")

sociald_att<-effects

load("Treatment Effects for Substantive Ideology Change.Rdata")

subideo_att<-effects

load("Treatment Effects for Ideology Change.Rdata")

ideo_att<-effects

load("Treatment Effects for Change No Opinion Leaders Followed.Rdata")

ol_follow_att<-effects

load("Treatment Effects for Change in Strength of Echo Chamber.Rdata")

echo_att<-effects

thermometer_att$variable<-"thermometer"
sociald_att$variable<-"sociald"
subideo_att$variable<-"subideo"
ideo_att$variable<-"ideo"
ol_follow_att$variable<-"olfollow"
echo_att$variable<-"echo"

final_att<-rbind(thermometer_att,
                 sociald_att,
                 subideo_att,
                 ideo_att,
```

```
                ol_follow_att,
                echo_att)

final_att$variable<-factor(final_att$variable,
                           levels=c(
                            "echo",
                            "olfollow",
                            "subideo",
                            "ideo",
                            "sociald",
                            "thermometer"
                           ),
                           labels=c(
                          "Change in % Co-Partisans Followed on Twitter",
                          "Change in # of Political Accounts Followed",
                          "Change in Liberalism/Conservativism Index",
                          "Change in 7-Point Self-rated Ideology",
                          "Change in Desired Social Distance from Opposing Party",
                          "Change in Opposing Party Feeling Thermometer"))

#create plot
library(ggplot2)
p<-ggplot(final_att)+
  geom_hline(yintercept = 0, colour = gray(1/2), lty = 2)+

  geom_linerange(aes(x = variable, ymin = ATT.lb,
                     ymax = ATT.ub),
                 lwd = 1, position = position_dodge(width = 1/2),
                 colour="skyblue")+

    geom_point(aes(x=variable, y=ATT),
               position = position_dodge(width = 1/2),
               size=2, colour="#71084D")+
  theme(axis.text=element_text(size=10, face="bold",colour="black"),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        panel.background=element_blank(),
        axis.title=element_text(size=12, colour="black"),
        legend.position="none",
        legend.key = element_blank(),
        legend.title=element_blank())+
  labs(x="",y="")+
  coord_flip()

 ggsave(p, file="ATT_Plot.eps", width=10, height=6, device = "eps", dpi = 300)

#now create plots for heterogeneous treatment effects reported in main text
#additional heterogeneous treatment effects reported in another section below

therm_tau_post<-thermometer$tau
therm_tauhat<-colMeans(therm_tau_post)
imputed_data$therm_effects<-therm_tauhat
```

```
sociald_tau_post<-sociald$tau
sociald_tauhat<-colMeans(sociald_tau_post)
imputed_data$sociald_effects<-sociald_tauhat

subideo_tau_post<-subideo$tau
subideo_tauhat<-colMeans(subideo_tau_post)
imputed_data$subideo_effects<-subideo_tauhat

ideo_tau_post<-ideo$tau
ideo_tauhat<-colMeans(ideo_tau_post)
imputed_data$ideo_effects<-ideo_tauhat

oplead_tau_post<-num_oplead_follows$tau
oplead_tauhat<-colMeans(oplead_tau_post)
imputed_data$oplead_effects<-oplead_tauhat

echo_tau_post<-echo_change$tau
echo_tauhat<-colMeans(echo_tau_post)
imputed_data$echo_effects<-echo_tauhat

#news interest plot

library(viridis)

forggplot<-imputed_data[,c(24:29,22)]

plotter<-forggplot %>%
  gather(Variable, Value,  -high_news_interest)

plotter$high_news_interest[plotter$high_news_interest==0]<-
  "Low News Interest"
plotter$high_news_interest[plotter$high_news_interest==1]<-
  "High News Interest"

plotter$Variable<-factor(plotter$Variable,
                         levels=c(
                        "therm_effects",
                        "sociald_effects",
                        "ideo_effects" ,
                        "subideo_effects",
                        "oplead_effects",
                        "echo_effects"
                          ),
                         labels=c(
                        "Change in Thermometer Rating of Opposing Party",
                        "Change in Desired Social Distance from Opposing Party",
                        "Change in Seven-Point Ideology Scale",
                        "Change in Liberalism/Conservatism Index",
                        "Change in # Political Accounts Followed",
                        "Change in % Co-Partisans in Twitter Network"
                          )
                         )
```

```
newsintplot<-ggplot(plotter, aes(y=Value,
                                 x=as.factor(high_news_interest),
                                 fill=Variable))+
  #geom_jitter(size=.3)+
  geom_violin(trim=FALSE)+
  scale_fill_viridis(discrete=TRUE)+
  #geom_boxplot(outlier.shape = NA, fill = "skyblue")+
  facet_wrap(~Variable, scales="free", ncol=2)+
  theme_minimal()+
  theme(axis.title.x=element_blank(),
        axis.title.y=element_blank(),
        axis.text.x = element_text(size=8),
        axis.text.y = element_text(size=8),
        legend.position="none",
        strip.text = element_text(size = 8, face="bold"),
        plot.title = element_text(hjust = 0.5, face="bold"))

ggsave(newsintplot, file="HTE_News_Interest.eps",
                          device = "eps", dpi = 300, width = 8, height = 10)

#Twitter Frequency
forggplot<-imputed_data[,c(24:29,21)]

plotter<-forggplot %>%
  gather(Variable, Value,  -frequent_twitter_user)

plotter$frequent_twitter_user[plotter$frequent_twitter_user==0]<-
  "Infrequent Twitter Users"
plotter$frequent_twitter_user[plotter$frequent_twitter_user==1]<-
  "Frequent Twitter Users"

plotter$Variable<-factor(plotter$Variable,
                         levels=c(
                        "therm_effects",
                        "sociald_effects",
                        "ideo_effects" ,
                        "subideo_effects",
                        "oplead_effects",
                        "echo_effects"
                          ),
                         labels=c(
                        "Change in Thermometer Rating of Opposing Party",
                        "Change in Desired Social Distance from Opposing Party",
                        "Change in Seven-Point Ideology Scale",
                        "Change in Liberalism/Conservatism Index",
                        "Change in # Political Accounts Followed",
                        "Change in % Co-Partisans in Twitter Network"
                          )
                         )

twitterfreqplot<-ggplot(plotter, aes(y=Value,
                                     x=as.factor(frequent_twitter_user),
                                    fill=Variable))+
```

```
  #geom_jitter(size=.3)+
  geom_violin(trim=FALSE)+
  #geom_boxplot(outlier.shape = NA)+
  facet_wrap(~Variable, scales="free", ncol=2)+
  theme_minimal()+
 scale_fill_viridis(discrete=TRUE)+
  theme(axis.title.x=element_blank(),
        axis.title.y=element_blank(),
        axis.text.x = element_text(size=8),
        axis.text.y = element_text(size=8),
        legend.position="none",
        strip.text = element_text(size = 8, face="bold"),
        plot.title = element_text(hjust = 0.5, face="bold"))

ggsave(twitterfreqplot, file="HTE_Twitter_Frequency.eps",
       device = "eps", dpi = 300, width = 8, height = 10)


#Party Identification

forggplot<-imputed_data[,c(2,24:29)]

plotter<-forggplot %>%
  gather(Variable, Value,  -republican_wave_1)

plotter$republican_wave_1[plotter$republican_wave_1==0]<-
  "Democrats"
plotter$republican_wave_1[plotter$republican_wave_1==1]<-
  "Republicans"

plotter$Variable<-factor(plotter$Variable,
                         levels=c(
                        "therm_effects",
                        "sociald_effects",
                        "ideo_effects" ,
                        "subideo_effects",
                        "oplead_effects",
                        "echo_effects"
                         ),
                         labels=c(
                        "Change in Thermometer Rating of Opposing Party",
                        "Change in Desired Social Distance from Opposing Party",
                        "Change in Seven-Point Ideology Scale",
                        "Change in Liberalism/Conservatism Index",
                        "Change in # Political Accounts Followed",
                        "Change in % Co-Partisans in Twitter Network"
                         )
                         )

republicanplot<-ggplot(plotter, aes(y=Value,
                                    x=as.factor(republican_wave_1),
                                    fill=Variable))+
  #geom_jitter(size=.3)+
```

```
  geom_violin(trim=FALSE)+
  #geom_boxplot(outlier.shape = NA)+
  facet_wrap(~Variable, scales="free", ncol=2)+
  theme_minimal()+
  scale_fill_viridis(discrete=TRUE)+
  theme(axis.title.x=element_blank(),
        axis.title.y=element_blank(),
        axis.text.x = element_text(size=8),
        axis.text.y = element_text(size=8),
        legend.position="none",
        strip.text = element_text(size = 8, face="bold"),
        plot.title = element_text(hjust = 0.5, face="bold"))

ggsave(republicanplot, file="HTE_Republican.eps", device = "eps", dpi = 300,
       width = 8, height = 10)
```

Sensitivity Analysis of IRA Exposure via Retweets

In the main text of our article, we provided two measures of IRA account interaction. We defined interactions as cases where respondents retweeted, mentioned, liked, or followed IRA accounts as well as those who may have been exposed to IRA accounts when they were mentioned in tweets by the people whom the respondent follows. We noted, in our discussion, that this last type of interaction does not include retweets of IRA accounts by people whom our respondents follow. In addition, we noted that trolls ocassionally retweet content by non-trolls, and we are unable– in some cases– to determine whether our respondents were exposed to such content via trolls or other parties, because a limitation of the way data are prepared by Twitter's Application Programming Interface (which we used to collect Twitter data from our respondents).

In the absence of more fine-grained exposure data, we present a sensitivity analysis below. In this analysis, we assume there is a linear relationship between our measure of echo chamber strength and the likelihood that respondents retweet trolls. This assumption is partially supported by Figure 1 in the main text of our article, which showed that the percentage of co-partisans followed on Twitter is the strongest predictor of IRA account engagement. Next, we calculated the echo chamber strength (alpha) of every person who was followed by one of the respondents in our study, using the ideological scoring techniques described above. We then recode our treatment indicator according to different thresholds of the echo chamber strength measure, rerun the propensity score matching, and re-estimate the Bayesian Causal Forest models for each of our outcomes. As Figure S4 below shows, we identified no significant impacts of IRA interaction for ten different values of alpha on each of the six outcomes analyzed in the main text of our paper.

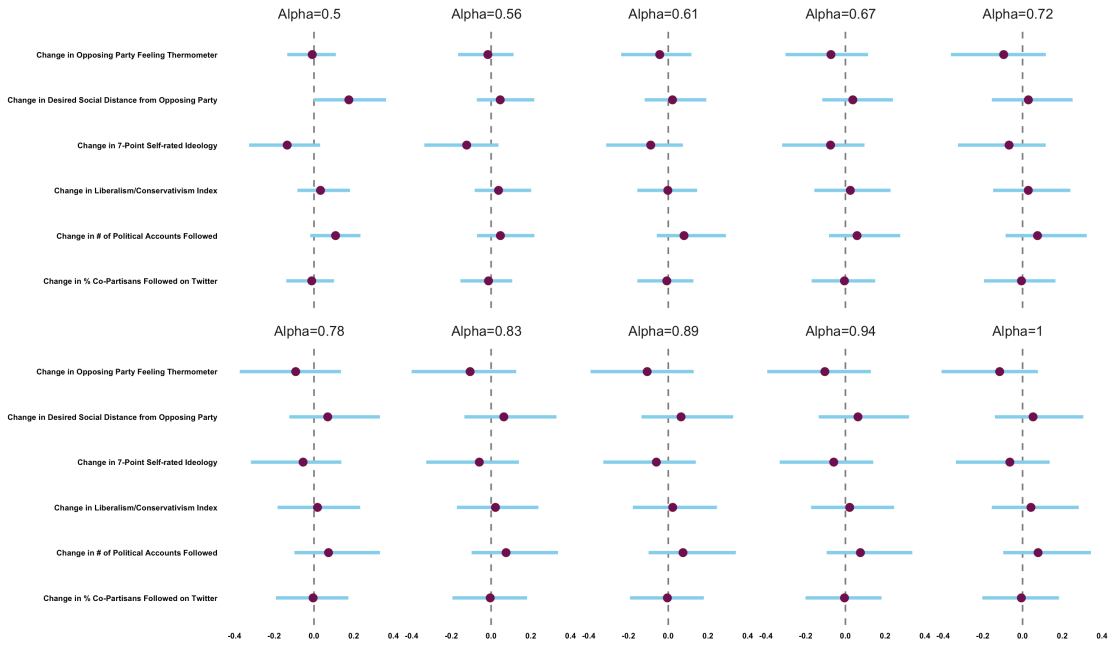

Figure S 4: Models from Main Text Re-Estimated with Synthetic IRA Retweet Exposure Measure

Models Including Pretreated Individuals

Because so many of our respondents who ever interacted with trolls did so before the study period, we also analyzed the entire sample of those who ever interacted with trolls. For these models, a dummy indicator of pretreatment status was included to attempt to control for the pre-study interaction with trolls. Results from these models, displayed in Figure S5, were insignificant across the six outcomes. In Table S3 we provide a breakdown of the number of respondents, including those who were pre-treated, who had each type of troll interaction.

```
load("For Troll Influence Models.Rdata")

#ONLY STUDY PERIOD
#getting time of interaction info
preincl <- troll_analysis_data %>% filter(ind_intbefore == 1 | ind_intduring == 1)

alltroll_pre <- trolltiming %>%
  filter(caseid %in% preincl$caseid) %>%
  mutate(pdstudy = ifelse(date < date_w2, 1, 0)) %>%
  group_by(caseid) %>%
  summarise(count = sum(pdstudy)) %>%
  select(caseid, count)

#getting type of interaction info
alltrolltype_pd <- trolltiming %>%
  filter(caseid %in% preincl$caseid) %>%
                      group_by(caseid) %>%
                      filter(date < date_w2) %>%
                      ungroup() %>%
                      spread(type, date)
#fix name issue
```

```r
library(janitor)


alltrolltype_pd <- clean_names(alltrolltype_pd)

alltrolltype_pd <- alltrolltype_pd %>%
  mutate(follow = ifelse(is.na(troll_follow), 0, 1),
         mention = ifelse(is.na(mention_trolls), 0, 1),
         favorite = ifelse(is.na(troll_favorite), 0, 1),
         friendmention = ifelse(is.na(troll_mention_by_friend), 0, 1)) %>%
  group_by(caseid) %>%
  summarise(numfollow = sum(follow),
            nummention = sum(mention),
            numfavorite = sum(favorite),
            numfriendmention = sum(friendmention),
            numengage = sum(nummention, numfavorite),
            numexposure = sum(numengage, numfollow, numfriendmention),
            follow = ifelse(numfollow > 0, 1, 0),
            mention = ifelse(nummention > 0, 1, 0),
            favorite = ifelse(numfavorite > 0, 1, 0),
            friendmention = ifelse(numfriendmention > 0, 1, 0))

#merge info with entire data
troll_analysis_data_pd <-
  left_join(troll_analysis_data, alltroll_pre, by = "caseid")
troll_analysis_data_pd <-
  left_join(troll_analysis_data_pd, alltrolltype_pd, by = "caseid")

troll_analysis_data_pd <- troll_analysis_data_pd %>%
  mutate(numfollow = replace_na(numfollow, 0),
         nummention = replace_na(nummention, 0),
         numfavorite = replace_na(numfavorite, 0),
         numfriendmention = replace_na(numfriendmention, 0),
         follow = replace_na(follow, 0),
         mention = replace_na(mention, 0),
         favorite = replace_na(favorite, 0),
         friendmention = replace_na(friendmention, 0),
         engage = ifelse(mention != 0 | favorite != 0, 1, 0),
         exposure = ifelse(mention != 0 | favorite != 0 | follow != 0 |
                             friendmention != 0, 1, 0))

sumtable <- troll_analysis_data_pd %>%
  summarise(totalmention = sum(mention),
            percmention = totalmention/1239*100,
            totalfavorite = sum(favorite),
            percfavorite = totalfavorite/1239*100,
            totalengage = sum(engage),
            percengage = totalengage/1239*100,
            totalfollow = sum(follow),
            percfollow = totalfollow/1239*100,
            totalfm = sum(friendmention),
            percfm = totalfm/1239*100,
            totalexpose = sum(exposure),
```

```
            percexpose = totalexpose/1239*100
  )

totaln <- c(sumtable$totalengage, sumtable$totalexpose,
            sumtable$totalmention,
            sumtable$totalfavorite, 0,
            sumtable$totalfollow, sumtable$totalfm)

perc <- c(sumtable$percengage, sumtable$percexpose,
          sumtable$percmention,
          sumtable$percfavorite, 0,
          sumtable$percfollow, sumtable$percfm)

sumtablekable3 <- cbind(totaln, perc)
rownames(sumtablekable3) <- c("Any Direct Engagement", "Any Interaction",
                              "Ever Mention Troll", "Ever Like Troll",
                              "Ever Like Troll Mention",
                             "Ever Follow Troll", "Ever Friend Mention")

colnames(sumtablekable3) <- c("Number", "Percent")

kable(sumtablekable3,
      caption = "Troll Interaction Summary Statistics - Pretreat Models",
      digits = 2, format = "latex") %>%
   row_spec(1:2, bold = T) %>%
  kableExtra::kable_styling(latex_options = "hold_position")
```

Table S 3: Troll Interaction Summary Statistics - Pretreat Models

	Number	Percent
Any Direct Engagement	**81**	**6.54**
Any Interaction	**110**	**8.88**
Ever Mention Troll	56	4.52
Ever Like Troll	33	2.66
Ever Like Troll Mention	0	0.00
Ever Follow Troll	9	0.73
Ever Friend Mention	34	2.74

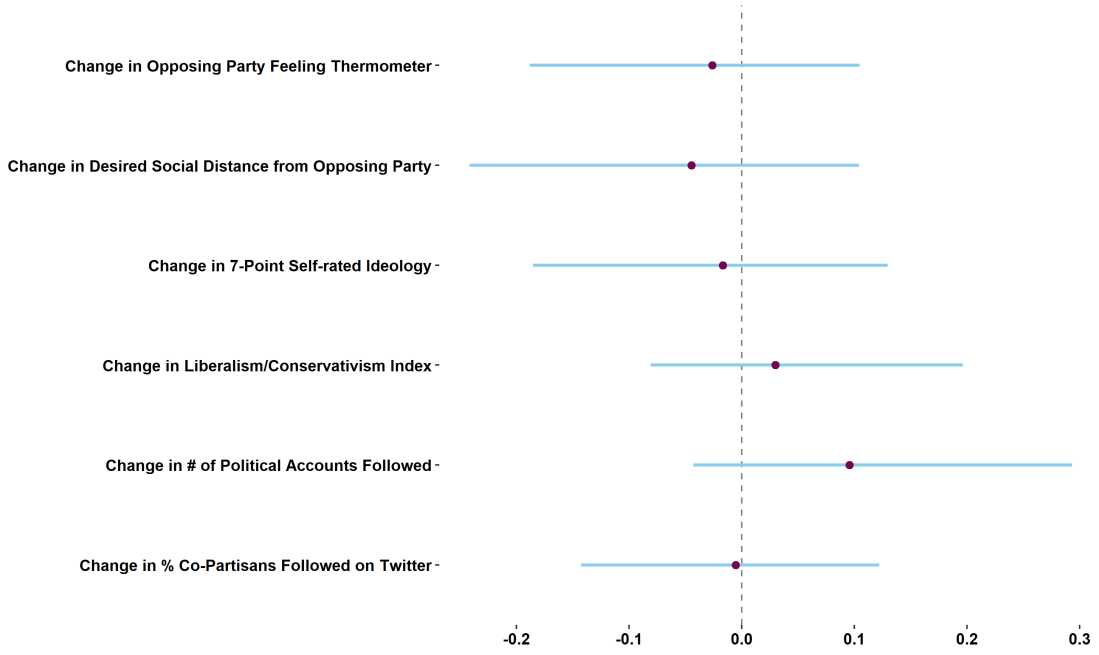

Figure S 5: Models Including Pretreated Individuals

Additional Analysis of Dosage Effects

We also expanded our analysis to include troll interactions that occured before and after the initial one-month period between survey waves. We obtained data collected about our respondents by the YouGov survey firm between February 2016 and April 2018. Unfortunately, the only relevant survey question consistently available across this time period was the seven-point liberal/conservative scale question, which we were able to obtain for 1,237 study respondents. We defined treatment as interacting with a troll between the earliest and latest measure of ideology available for each respondent in the YouGov profile dataset. Following the previous analysis, we exclude pre-treated respondents who interacted with trolls prior to the first their first measure of ideology. We transformed each measure of ideology onto a 0-1 scale, calculated the difference, and then standardized that difference to have a mean of 0 and a standard deviation of 1.

We ran separate models for different levels of dosage—with the treated group in each model defined as respondents who interacted with trolls one or more times (Treated = 213, Control = 1,024), two or more times (Treated = 110, Control = 1,127), or three or more times (Treated = 67, Control = 1,170). We ran models with the same specification previously described to understand whether these interactions influenced ideology. Additionally, we ran separate models in which we operationalized treated individuals as only those who had direct enagements with trolls (mentioning a troll, liking a troll's tweet or a tweet that mentions a troll)— since it is possible that activities included in our broader operationalization of troll interactions (e.g., following a troll) are insufficient to change political attitudes.

Figure 6 in the main text of our manuscript contains the results from 6 separate models, with different types of troll interactions (exposure, engagement only) and different number of interactions (1 or more, 2 or more, 3 or more). In Table S4 (below) we specify the number of respondents in the treatment and control groups for each model.

We further considered dosage effects by using the interaction variable as a count measure with levels of 1 or more, 2 or more, and 3 or more interactions in OLS models for the six outcomes. Figure S6 (in this document, below) indicates there were no significant effects for differing dosage levels of all engagements with trolls on Twitter for all six outcomes we studied.

Table S 4: Number of Respondents by Treatment Group

Interaction Type	Number of Interactions	N: Control Group	N: Treatment Group
All Engagements	1	1017	213
Direct Engagements Only	1	1103	127
All Engagements	2	1120	110
Direct Engagements Only	2	1173	57
All Engagements	3	1163	67
Direct Engagements Only	3	1199	31

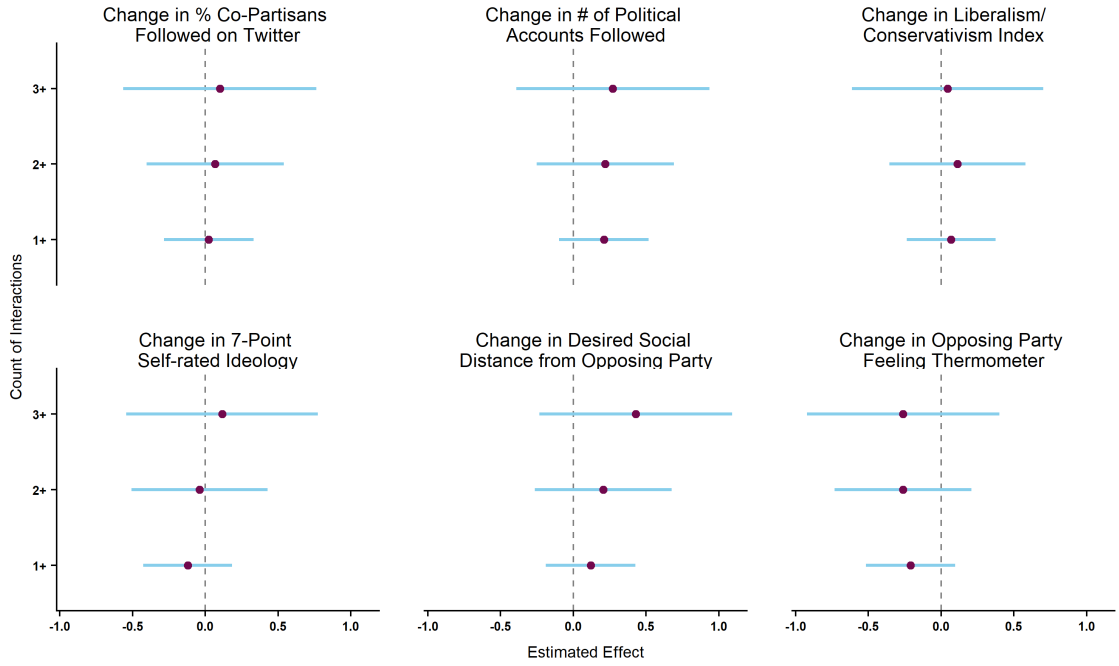

Figure S 6: Models Using Interaction with Trolls as a Count Variable

The Nature of Troll Interactions

In the main text of our manuscript, we noded that direct engagment with trolls may produce different effects than other forms of interaction on political attitudes and behaviors. Figure S7 reproduces the main model from our paper using only direct engagements. For convenience, Figure S8 presents the model that describes all forms of interaction (either direct or indirect) from the main text of our paper.

Figures S9 and S10 add dosage to these operationalizations. For direct engagements, we defined treated alternatively as one or more engagments and two or more engagements, and for the model that descirbes all forms of troll interactions we compared one or more, two or more, and three or more interaction dosage levels. One important note for these dosage models is that as the definition of "treated" grew smaller, it was more difficult to do propensity score matching with the amount of covariates we used in the main models in a paper—since employing all covariates resulted in perfect separation between treatment and control groups. Because of this, for the propensity score matching step, we dropped several covariates. Specifically, for the direct engagement model at dosage level of two or more engagements, the covariates Republican, frequent Twitter user, high news interest, and Western region were dropped—but only for the propensity score matching step. These covariates were still included in the Bayesian Causal Forest model. Similarly, for the interaction model of a dosage level of 3 or more interactions, the covariates of frequent Twitter user and high news interest were dropped from the propensity score portion of the analysis. Across each of these models, the effect of interacting with trolls remains insignificant for all of the outcome measures of political

beliefs and attitudes.

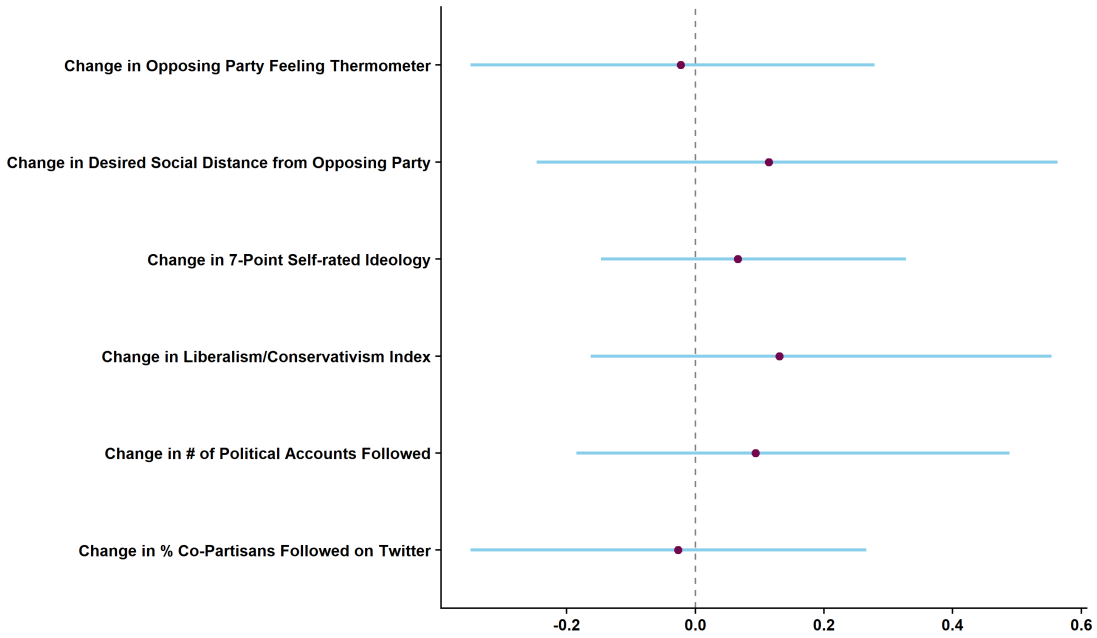

Figure S 7: Models that Examine Direct Engagement with Trolls: Treated Individuals Liked, Retweeted, Mentioned, or Liked a Mention of a Troll at Least Once

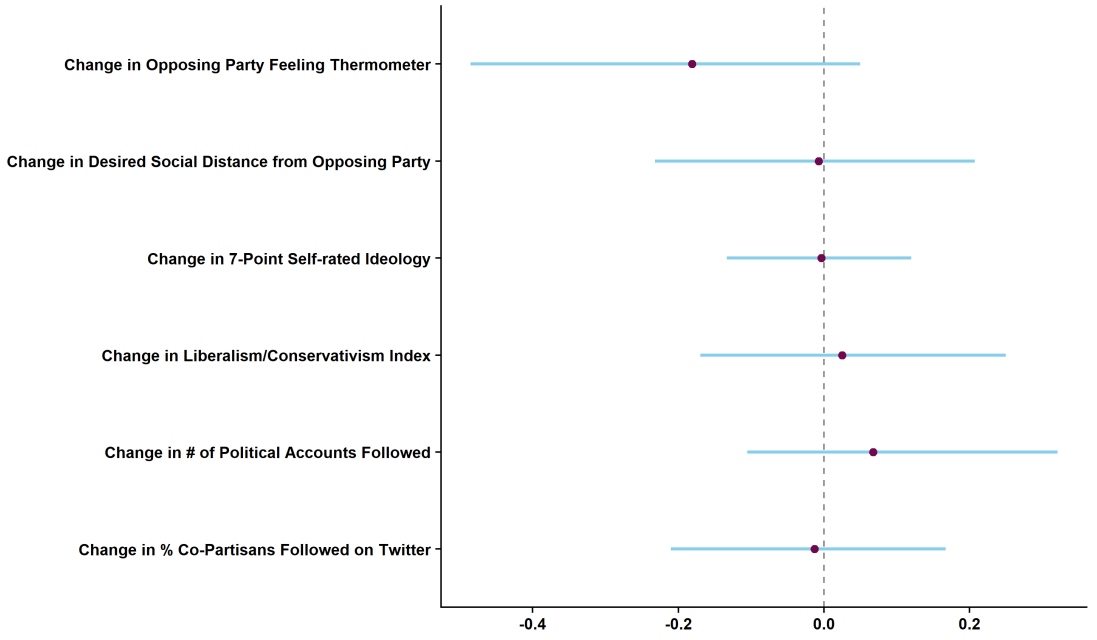

Figure S 8: Models that Examine All forms of Troll Interaction: Treated Individuals Liked, Retweeted, Mentioned, or Liked a Mention on a Troll at Least once—or Followed a Troll, or Had a Friend Mention a Troll at Least Once

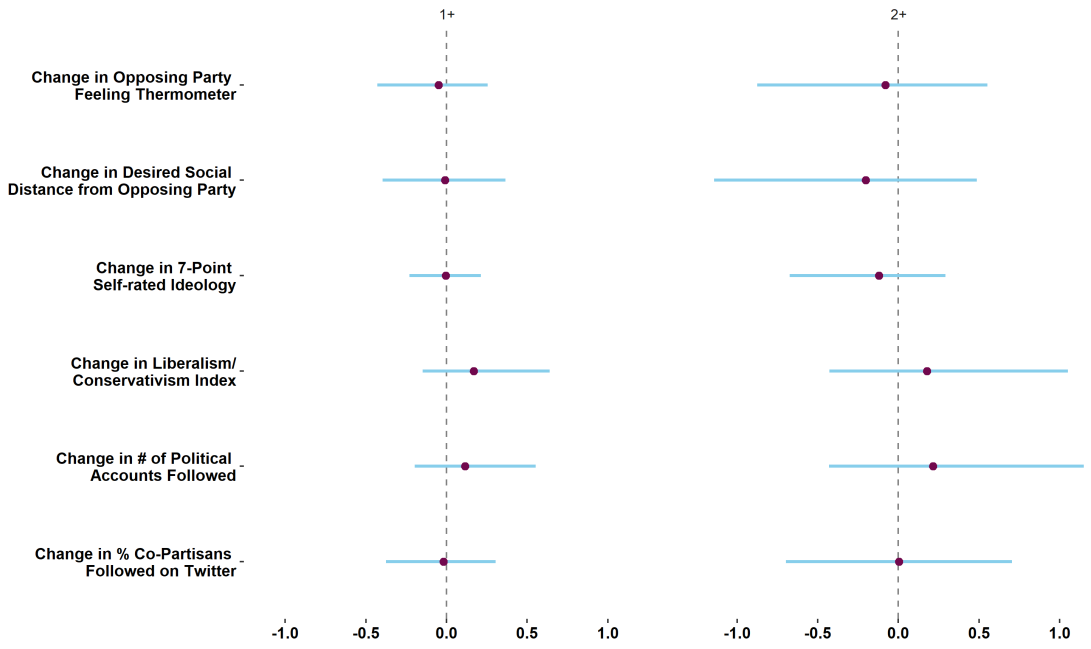

Figure S 9: Direct Engagement Models Testing Dosage Effects

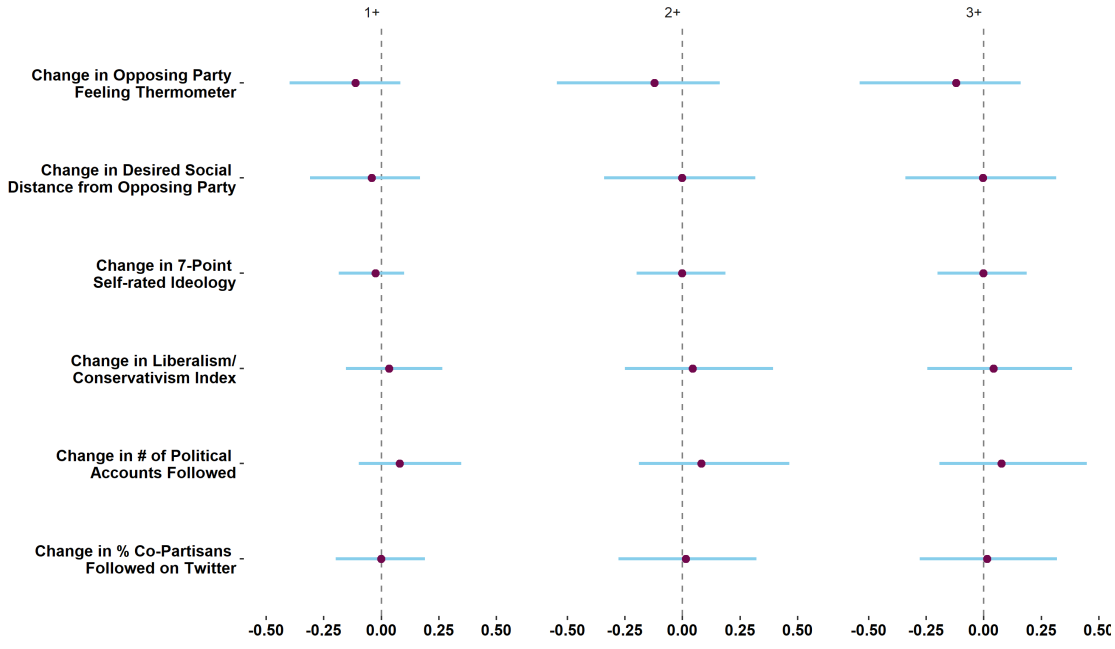

Figure S 10: Models that Examine Dosage Effects for All Forms of Troll Interaction

Additional Analysis of Heterogeneous Treatment Effects

The figures that follow report the effects of IRA account interaction on each of our outcomes by all of the covariates in the models in the main text of the paper. Readers should be cautioned not to over-interpret

some of the differences across categorical variables below, as some of them are very sparsely populated by those who interacted with IRA accounts, and the overall rate of interaction was relatively low for many subgroups. For example, there was only one non-white respondent who interacted with IRA trolls according to our measure.

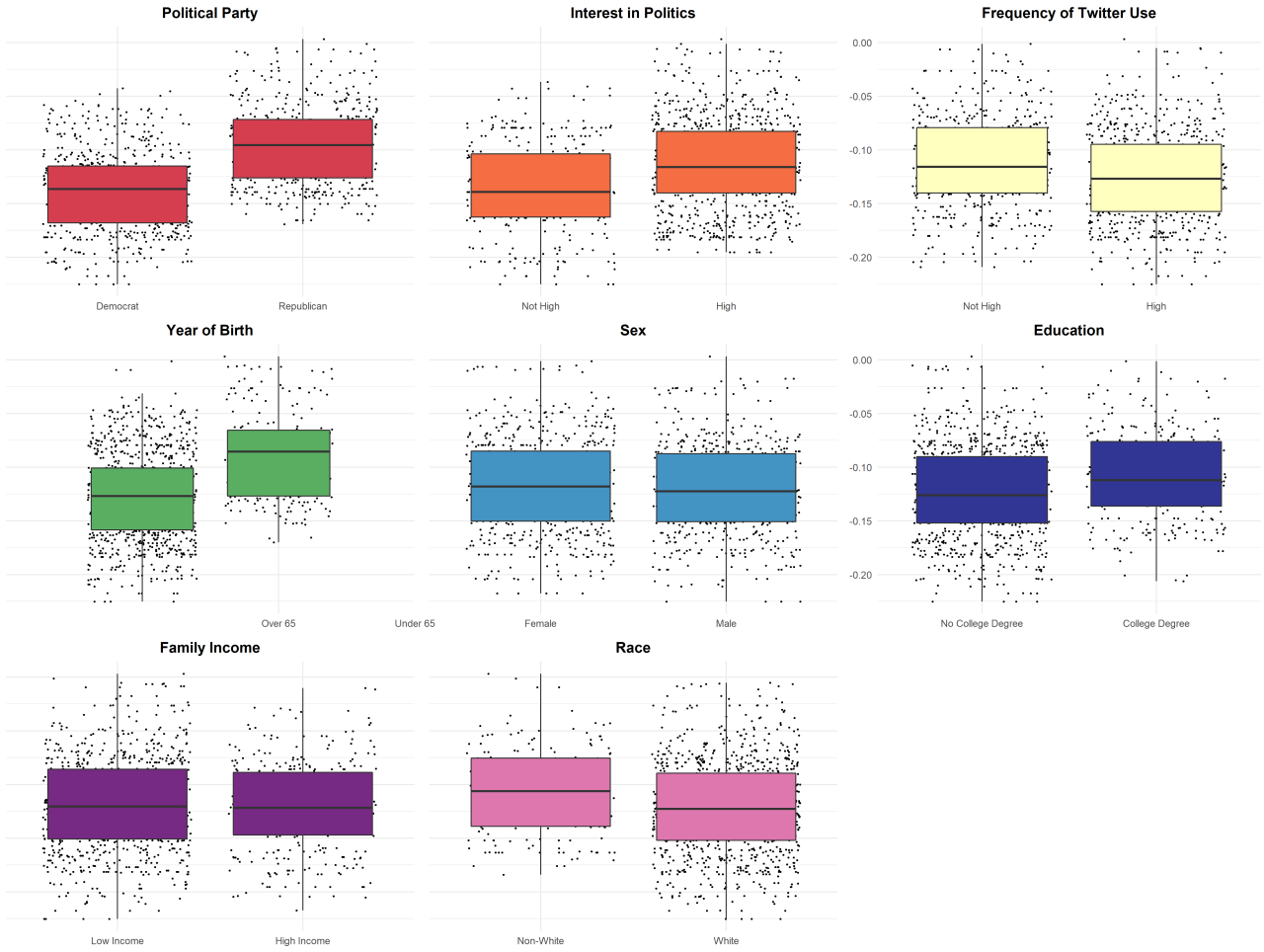

Figure S 11: Treatment Effects of Troll Engagement on Change in Thermometer Ratings of Opposing Political Party by Various Covariates

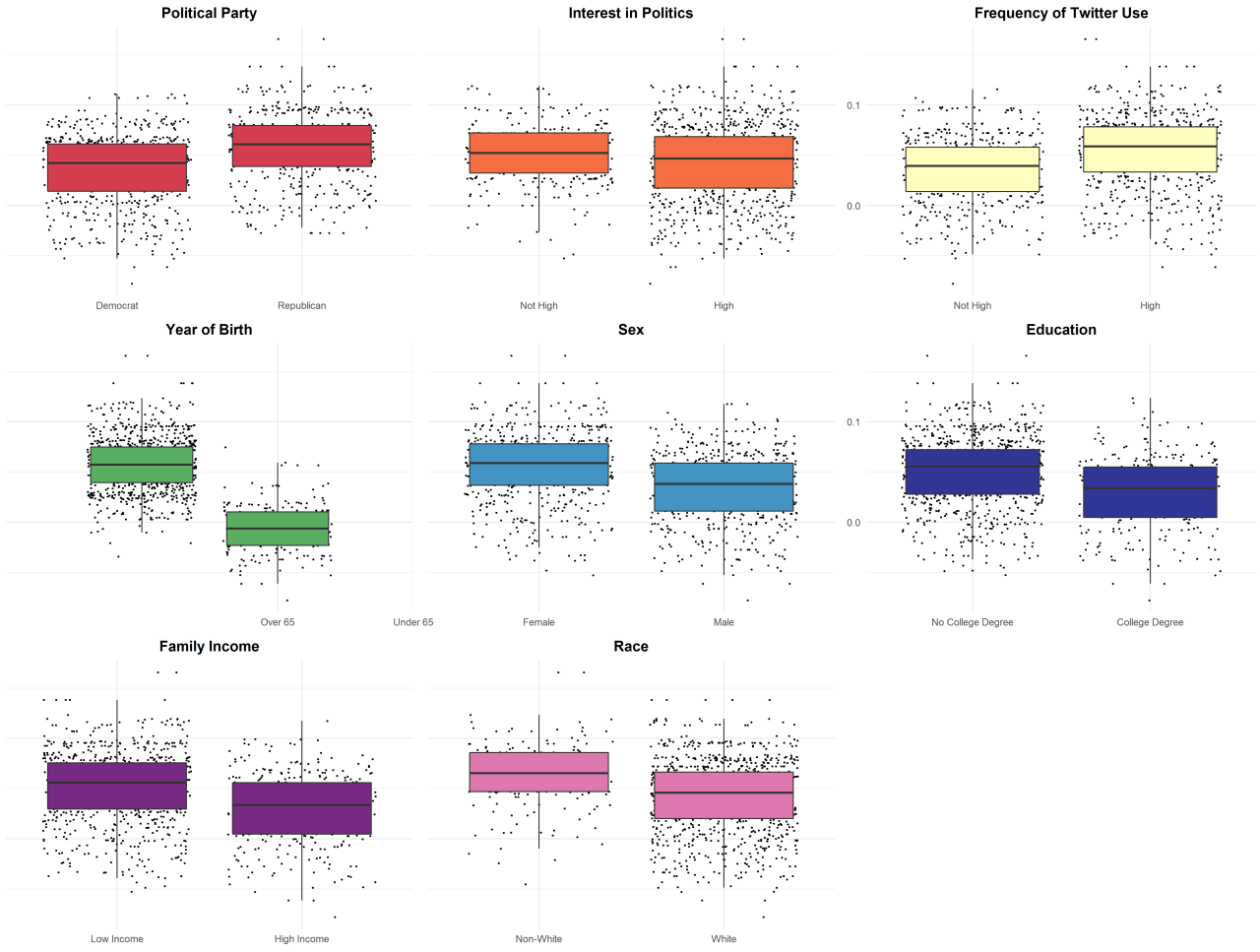

Figure S 12: Treatment Effects of Troll Engagement on Change in Desired Social Distance from Opposing Political Party by Various Covariates

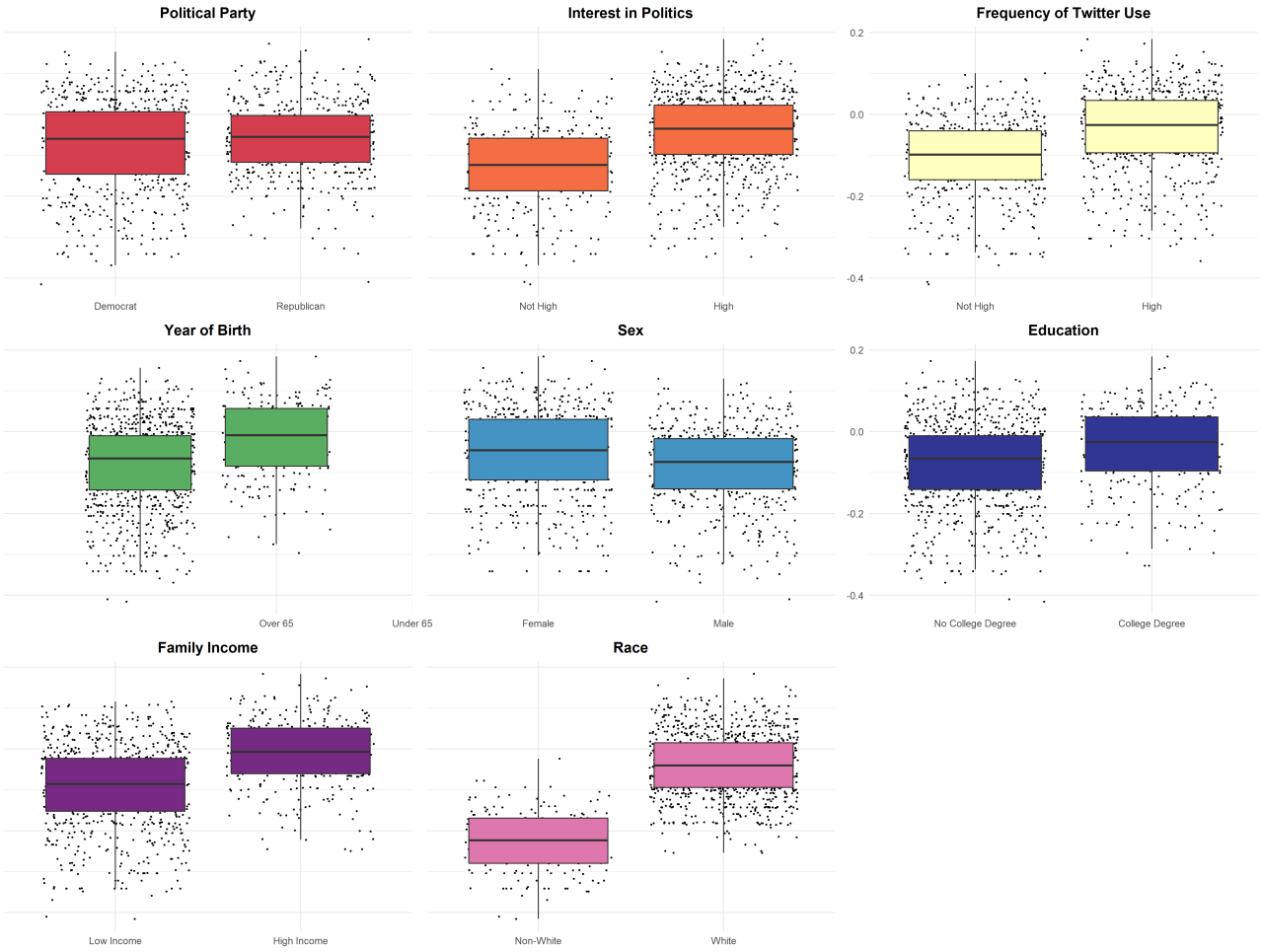

Figure S 13: Treatment Effects of Troll Engagement on Change in Seven-Point Ideology Scale by Various Covariates

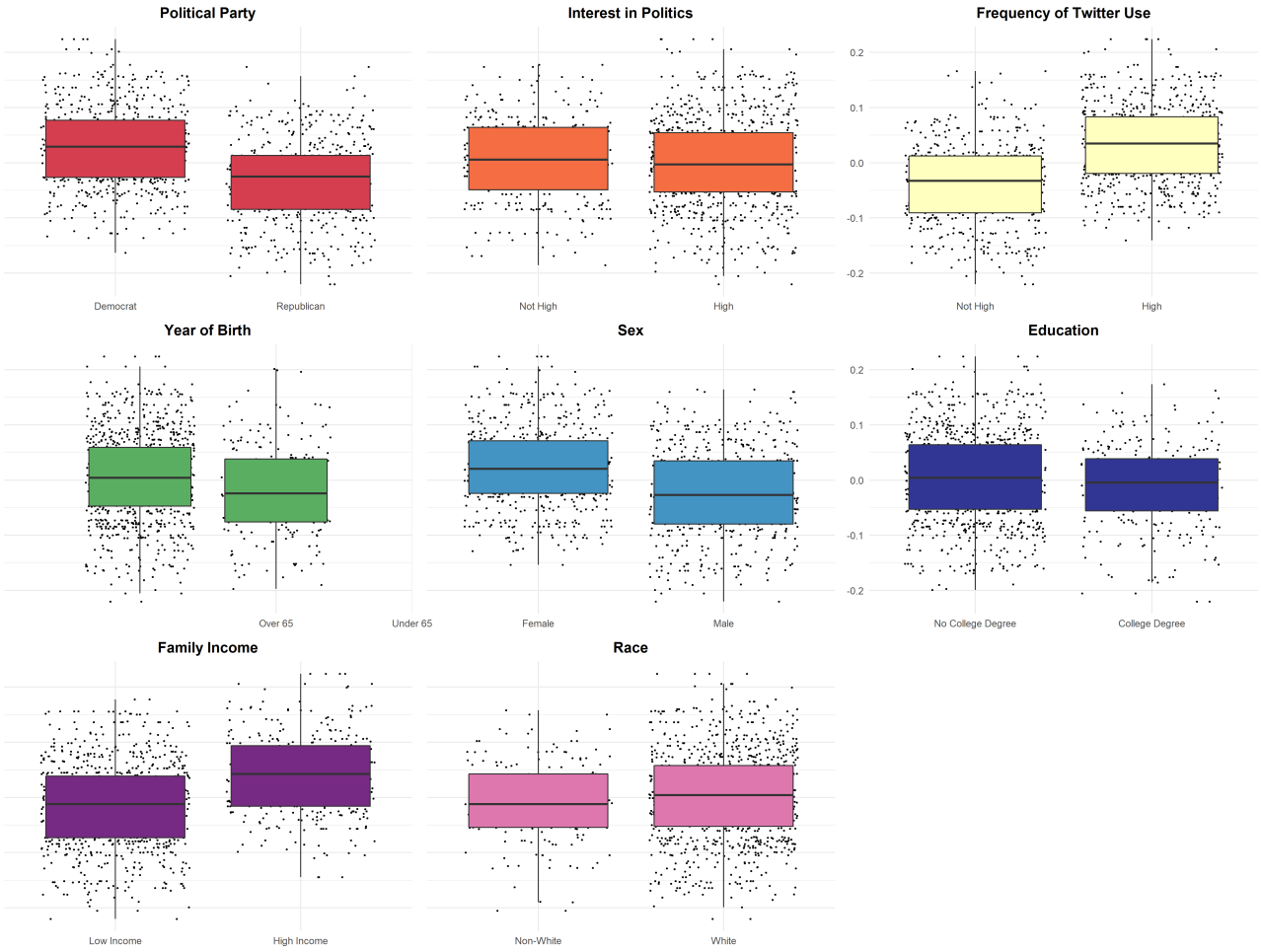

Figure S 14: Treatment Effects of Troll Engagement on Change in Liberal/Conservatism Index by Various Covariates

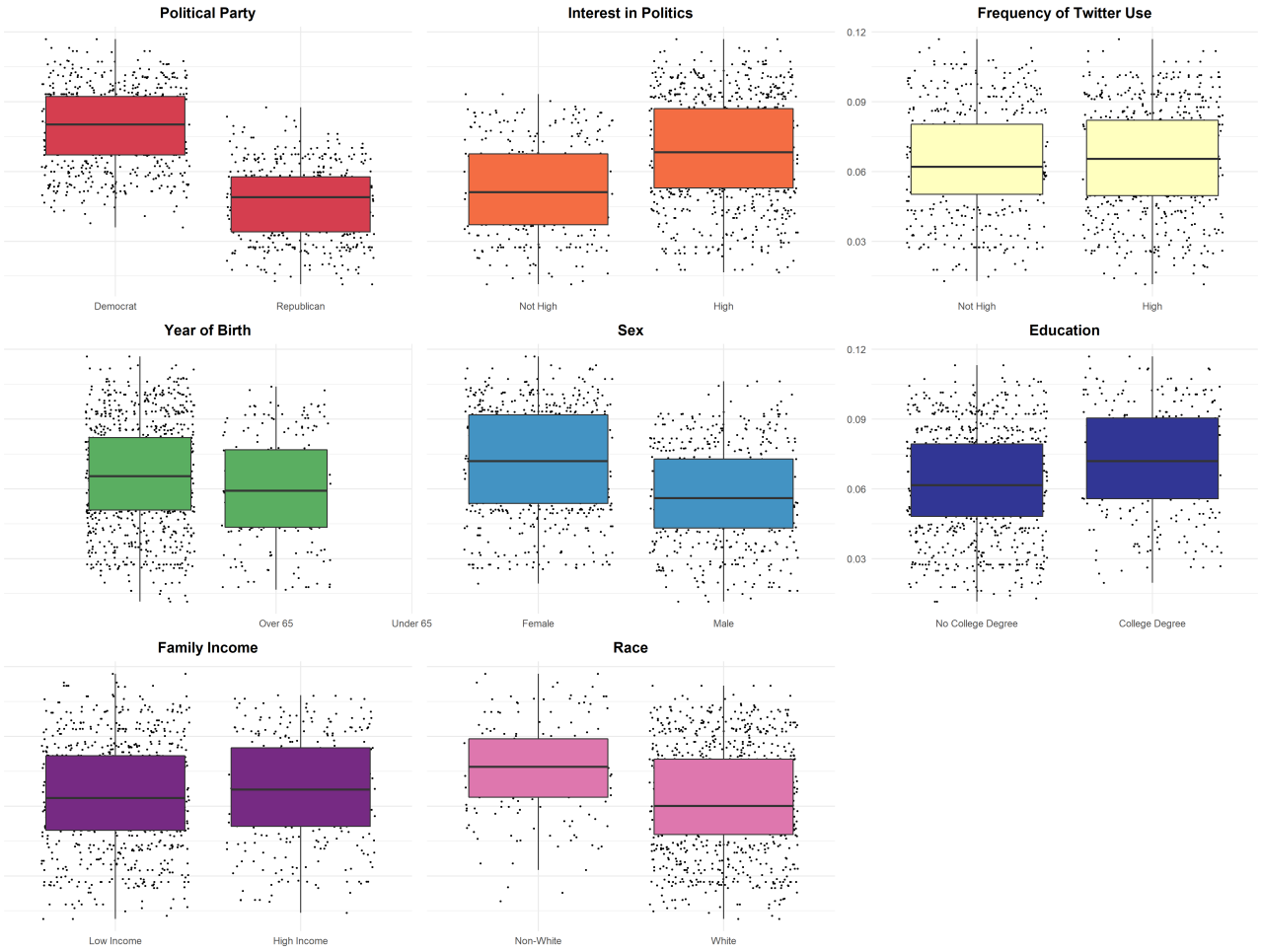

Figure S 15: Treatment Effects of Troll Engagement on Change in Number of Political Accounts Followed on Twitter by Various Covariates

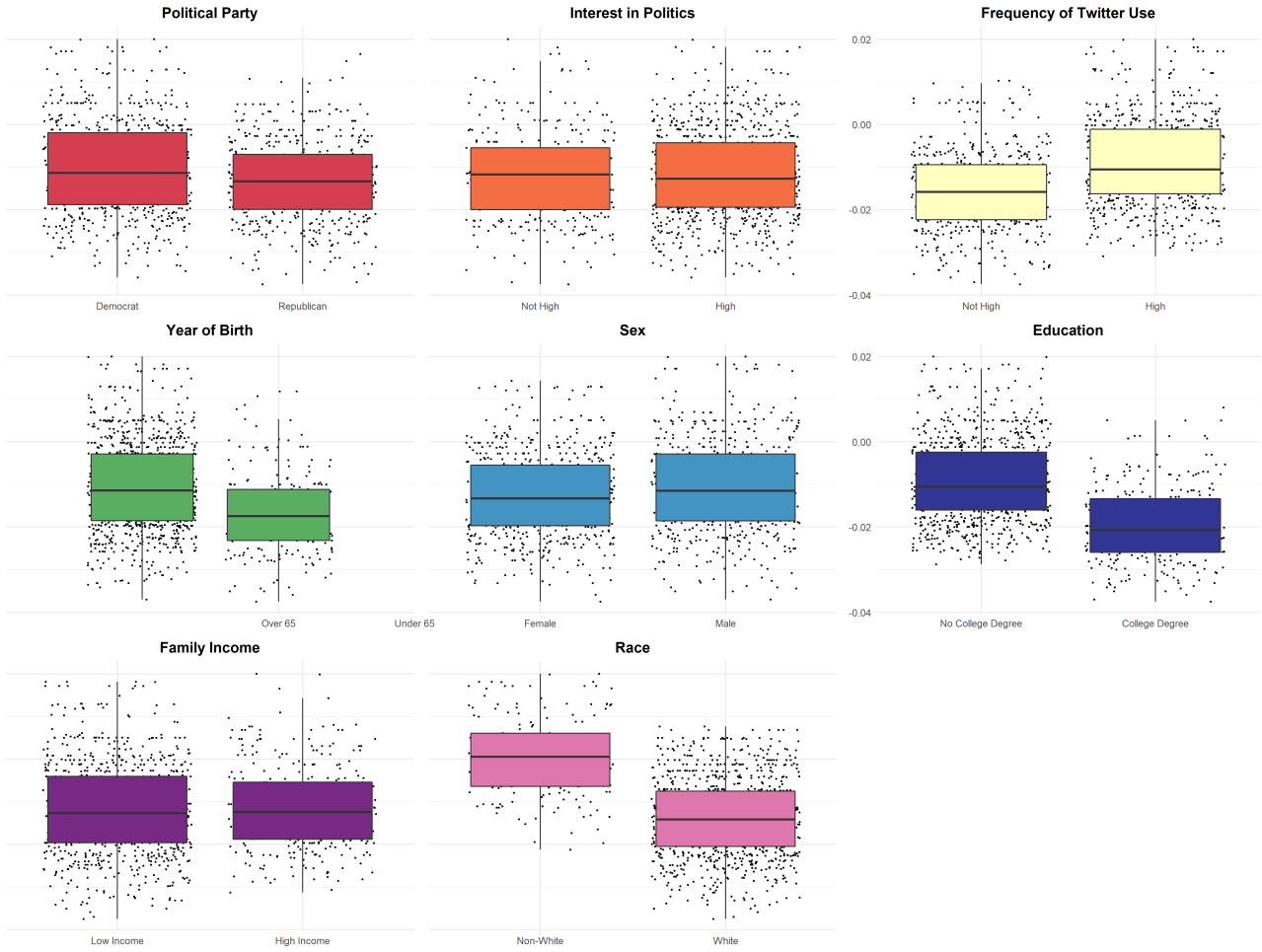

Figure S 16: Treatment Effects of Troll Engagement on Change in Percentage of Co-Partisans in Twitter Network by Various Covariates

Content Analysis of IRA Communications

In order to verify that respondents in our study were exposed to the type of social media content that might influence their political attitudes or behaviors, we completed several types of content analysis. First, two members of the research team independently coded subsections of interactions between respondents and troll messages in order to assess what type of information was being communicated by trolls to respondents. More specifically, our coders measured whether each tweet included political content or mentioned a political elite. Tweets were coded as political if they referenced current events or debates related to government, politicians, political parties, liberals/conservatives, the country or U.S. government. Tweets were codeded as referencing a political elite if they described people directly involved with the U.S. government, institutions tied to politics, and people in other significant positions of power performing political acts. Intercoder reliability was high: alpha = 0.89 for any political content and apha=0.86 for tweets that mention political elites. Of all IRA Twitter messages, we determined that 92.3% contained political content (56.8% explicitly mentioned an elected official or other political personality). Furthermore—of the respondents who engaged with troll content during the study period—94.6% engaged with a political tweet, which means that virtually all troll interactions experienced by respondents were political in nature.

Next, we identified the top 100 words in messages that were liked by respondents in our study. As Table S5 (below) shows, nearly all of these words have political connotations, and many concern the debate about an NFL football player kneeling during the national anthem, or the conflict between President Trump and Senator Elizabeth Warren. To further illustrate the type of content in the IRA interactions we studied, we provide several full-text examples. For example, our sample of troll interactions includes the following tweet from the "TEN_GOP" troll account——one of the most prolific and popular conservative accounts: "Father God we pray that this Trump Victory Tsunami will translate ALL ACROSS AMERICA!…LAND OF THE FREE! HOME OF THE BRAVE!" An example of a tweet from a popular liberal account, "wokeluisa", included: "Drunk white nationalist Steve Bannon is campaigning with pedophile Roy Moore tonight. I wonder if Moore asked Bannon to bring his daughters." The majority of troll content in which respondents engaged is similar to these example tweets: provocative, far-reaching, and intentionally political in nature; they appear aimed at polarizing political attitudes. In other words, the trolls included in our analysis are exactly the type of trolls that have been the issue of concern during and after the 2016 presidential election.

Additionally, several tweets had more than one interaction by our respondents. For example, this tweet by "wokeluisa" was retweeted by 10 of our respondents: "Don't ever tell me kneeling for the flag is disrespectful to our troops when Trump calls a sitting Senator 'Pocahontas' in front of Native American war heroes." Of tweets that had multiple interactions with respondents, many were political and also had a very large amount of favorites overall. A tweet by "kanijjackson" saying "This is how a United States President should look at Putin," was liked by 6 of our respondents and had 77,666 favorites in total.

Attrition Bias

The table below indicates there is no evidence of attrition bias by treatment condition. For additional tests for attrition bias see Bail et al. (2018).

Table S5: Attrition by Condition

Condition	Pre-Treatment Survey	Post-Treatment Survey
Control	495	413
Treatment	744	656

References

Bail, Christopher A., Lisa P. Argyle, Taylor W. Brown, John P. Bumpus, Haohan Chen, M. B. Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. 2018. "Exposure to Opposing Views on Social Media Can Increase Political Polarization." *Proceedings of the National Academy of Sciences* 115 (37): 9216. https://doi.org/10.1073/pnas.1804840115.

Barbera, Pablo. 2015. "Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data." *Political Analysis* 23 (1): 76–91. https://doi.org/10.1093/pan/mpu011.

Dimock, Michael, and Doherty Carroll. 2014. "Political Polarization in the American Public: How Increasing Ideological Uniformity and Partisan Antipathy Affect Politics, Compromise, and Everyday Life." *Pew Research Center Report*.

Fiorina, Morris P, Samuel J Abrams, and Jeremy Pope. 2006. *Culture War?: The Myth of a Polarized America*. Longman Publishing Group.

Hill, Seth J, and Chris Tausanovitch. 2015. "A Disconnect in Representation? Comparison of Trends in Congressional and Public Polarization." *The Journal of Politics* 77 (4): 1058–75.