Leveraging New Technologies and Interdisciplinarity to Study Political Behavior, Attitudes, and Beliefs

John Ternovski

Follow this and additional works at: https://elischolar.library.yale.edu/gsas_dissertations

Recommended Citation
Ternovski, John, "Leveraging New Technologies and Interdisciplinarity to Study Political Behavior, Attitudes, and Beliefs" (2021). Yale Graduate School of Arts and Sciences Dissertations. 428. https://elischolar.library.yale.edu/gsas_dissertations/428

This Dissertation is brought to you for free and open access by EliScholar – A Digital Platform for Scholarly Publishing at Yale. It has been accepted for inclusion in Yale Graduate School of Arts and Sciences Dissertations by an authorized administrator of EliScholar – A Digital Platform for Scholarly Publishing at Yale. For more information, please contact elischolar@yale.edu.
Abstract

Leveraging New Technologies and Interdisciplinarity to Study Political Behavior, Attitudes, and Beliefs

John Ternovski

2021

I make use of new technological and scholarly developments to study political sentiments and behavior in three independent papers. In my lead paper, I address an important consequence of political deepfakes (i.e., computer-manipulated video misinformation): does the provision of information about deepfakes cause people to disbelieve real political videos? Through a set of online survey experiments, I find that information that is typical of news coverage of deepfakes induces people to disbelieve real political information. My second paper uses new social media datasets to address pressing questions about how organized American far-right groups (e.g., neo-Nazis, white supremacists, etc.) recruit new members, and whether the rise of Trump was used as a catalyst in far-right recruitment efforts. I made use of prior sociological and anthropological research that found that far-right music scenes (featuring bands with such names as Aryan Terrorism) are a key part of day-to-day functioning of the overwhelming majority of far-right hate groups in the United States. As such, I made use of public databases of song listenership on the music social network, Last.fm, before and after Trump events. I find that online friends of frequent listeners of hate music were more likely to increase their levels of hate music listenership after Trump-related events (e.g., xenophobic tweets, primary election victories, etc.). Finally, in my third paper, I leverage new theoretical frameworks in the cognitive sciences and the growth of large-scale, data-driven voter mobilization programs among non-profit organizations to
address the puzzle of “voting habits.” Namely, prior research provides strong empirical
evidence that voting in one election makes the average individual more likely to vote in a
subsequent election, but this kind of turnout persistence does not comport with habit as it is
defined in psychological sciences (elections happen too infrequently and voting is never an
automatic behavior). So, in my third paper, I apply Duckworth and Gross’s (2020) Process
Model of Behavior Change to turnout persistence to bridge the gap between classic
economic models of voter turnout and the large body of rigorous empirical evidence
showing turnout persistence. I evaluate the concrete predictions made by this model in a
novel dataset of ~1.8 million voters across 9 different independent experiments.
Leveraging New Technologies and Interdisciplinarity to Study Political Behavior, Attitudes, and Beliefs

A Dissertation

Presented to the Faculty of the Graduate School

of

Yale University

in Candidacy for the Degree of

Doctor of Philosophy

By

John Ternovski

Dissertation Director: Peter M. Aronow

December 2021
# Contents

List of Figures ...................................................................................................................................................... iv
List of Tables ....................................................................................................................................................... vi

1. Introduction................................................................................................................................................. 1
   1.1: Does the Rise of Deepfakes Affect Political Accountability? ........................................................... 2
   1.2: Do Far-Right Hate Groups Use Newsworthy Trump-Related Events to Recruit Others? ........ 4
   1.3: What Motivates Americans to Continue to Turn Out to Vote? ...................................................... 6

2. The Negative Consequences of Informing Voters about Deepfakes: Evidence from Two Survey Experiments ........................................................................................................................................................ 12
   2.1: Significance Statement........................................................................................................................... 12
   2.2: Abstract.................................................................................................................................................... 12
   2.3: Introduction ............................................................................................................................................ 13
   2.4: Experimental Design............................................................................................................................. 17
      2.4.1: Study 1 - Fictional Politician......................................................................................................... 18
      2.4.2: Study 2 - Real Politician ................................................................................................................ 21
   2.5: Results and Discussion.......................................................................................................................... 23
      2.5.1: Heterogenous Treatment Effects ................................................................................................ 27
   2.6: Discussion and Conclusions ................................................................................................................ 28
   2.7: References ............................................................................................................................................... 29
   2.8: Supplementary Materials ....................................................................................................................... 35
      2.8.1: Deepfake Model Specifications.................................................................................................... 35
      2.8.2: Deviations from Pre-Analysis Plan (PAP) ................................................................................. 37
      2.8.3: Discernment of Deepfakes Without Deepfake Information .................................................. 38
      2.8.4: Additional Tables and Figures ..................................................................................................... 41
      2.8.5: Discussion of Secondary Research Question Results .............................................................. 54
      2.8.6: Discussion of Motivated Reasoning............................................................................................ 55
      2.8.7: Ethical Considerations .................................................................................................................. 56
      2.8.8: Materials ........................................................................................................................................... 57

3. The Impact of Trump Tweets and 2016 Victories on Hate Music Listenership ................................................. 105
   3.2: Introduction .......................................................................................................................................... 105
3.3: Motivation and Prior Research ................................................................. 107
  3.3.1: Hate Groups ....................................................................................... 109
  3.3.2: The Relationship between Hate Groups and Trump ......................... 111
3.4: Theory ..................................................................................................... 112
3.5: Methods and Materials ........................................................................ 115
3.6: Results .................................................................................................... 118
  3.6.1: Data Censoring and Amended Analysis ........................................... 118
  3.6.2: Sample .............................................................................................. 119
  3.6.3: Results .............................................................................................. 121
  3.6.4: Placebo Checks ................................................................................ 124
3.7: Discussion and Conclusions ................................................................. 127
3.8: References .............................................................................................. 129
3.9: Supplementary Materials .................................................................... 138
  3.9.1: Data and Materials ......................................................................... 138
  3.9.2: Issues with Naïve Sample of Top Listeners of Hate Music ............... 148
  3.9.3: Registered Analyses in Original PAP .............................................. 152
  3.9.4: Registered Analyses in Amended PAP ............................................. 154
  3.9.5: Robustness Checks ......................................................................... 157
  3.9.6: Placebo Checks ................................................................................ 161
  3.9.7: Original PAP .................................................................................... 167
  3.9.8: Amended PAP ................................................................................ 189
  4.1: Abstract .............................................................................................. 199
  4.2: Introduction .......................................................................................... 200
  4.3: The Calculus of Voting ....................................................................... 203
  4.4: Reconciling “Habit” with the Calculus of Voting .................................. 207
  4.5: Skipping the Calculus of Voting; Voting as a “Habit” ........................... 211
  4.6: “Voting-as-a-Habit” Literature and the PMBC Model ......................... 213
  4.7: Data ..................................................................................................... 220
  4.7.1: Predictions ....................................................................................... 228
4.8: Results.................................................................................................................................................... 229

4.8.1: Upstream Effects ......................................................................................................................... 230

4.8.2: Analysis Strategy ........................................................................................................................... 230

4.8.3: Overall Downstream Effects ..................................................................................................... 233

4.8.4: Evaluating Predictions from Prior Research ........................................................................... 234

4.9: Discussion and Conclusion ................................................................................................................ 241

4.10: References ........................................................................................................................................... 246

4.11: Supplementary Materials ................................................................................................................... 253

4.11.1: Additional Tables and Figures ................................................................................................. 253

4.11.2: Anonymized Canvasser Script and Training Documents.................................................... 255
List of Figures

Figure 2.1: Screencap of video used in Study 1........................................................................................................ 19

Figure 2.2: Information about deepfakes induced disbelief in accompanying video regardless of whether the video is real or fake............................................................................................................................................. 25

Figure 2.3: Voters were unable to discriminate between a real video and a deepfake (Study 1)......... 40

Figure 2.4: Study 2 - Impacts of information about deepfakes on disbelief (3-point scale):
unconvinced that politician believes what is being said................................................................................................. 43

Figure 2.5: Study 2 - Impacts of information about deepfakes on disbelief (binary): don’t know if
politician truly believes what is being said .......................................................................................................................... 44

Figure 2.6: Information about deepfakes had no impact on trust in future videos found on TV and online .................................................................................................................................................................................. 53

Figure 3.1: Friends of top hate listeners who are more active and more central have significant effects .................................................................................................................................................................................. 123

Figure 3.2: Impact of placebo events on hate music listenership of friends of far-right users......... 125

Figure 3.3: Impact of Trump-related events on hate music listenership of non-US friends of far-right users............................................................................................................................................................................................................. 126

Figure 3.4: Impact of Trump-related events on non-hate music listenership of friends of far-right users............................................................................................................................................................................................................... 127

Figure 3.5: Uncensored (left) and censored (right) artist pages................................................................................. 143

Figure 3.6: Uncensored (left) and censored (right) listener pages................................................................................. 143

Figure 3.7: Percent of daily listening that is hate music in our naïve sample of top listeners of hate music ............................................................................................................................................................................................................ 149

Figure 3.8: Percent of daily listening that is hate music among users who were top listeners of more hate bands than average ............................................................................................................................................................................................................. 150

Figure 3.9: Percent of daily listening that is hate music among users who were top listeners of the same hate band more times than average ........................................................................................................................................................................................................... 151
Figure 3.10: Percent of daily listening that is hate music among users who were top listeners of the same hate band more than average and more hate bands than average ................................................. 152

Figure 3.11: Friends of top hate listeners who are more active and more central have statistically significant Trump effects (Alternate 1) ........................................................................................................ 155

Figure 3.12: Friends of top hate listeners who are more active and more central have statistically significant Trump effects (Alternate 2) ........................................................................................................ 156

Figure 3.13: Attenuation of Trump effect on friends of far-right listeners across number of hate bands the far-right user was a top listener of .............................................................................................. 158

Figure 3.14: Attenuation of Trump effect on friends of far-right listeners across number of times the far-right user was a top listener of the same hate band ................................................................. 158

Figure 3.15: Attenuation of Trump Effect on friends of far-right listeners across far-right user’s eigenvector centrality in hate music listener network ................................................................. 159

Figure 3.16: Trump effect on friends of far-right listeners in terms of % of daily tracks that are hate music .................................................................................................................................................................. 160

Figure 3.17: Impact of Trump-related events on non-hate music listenership of friends of far-right users .................................................................................................................................................................. 162

Figure 3.18: Impact of placebo events on hate music listenership of friends of far-right users........ 165

Figure 3.19: Impact of Trump-related events on hate music listenership of non-US friends of far-right users .................................................................................................................................................................. 166

Figure 4.1: PMBC model as applied to voting before a voting “habit” (left) and after a voting “habit” develops (right) .................................................................................................................................................................. 212

Figure 4.2: Average treatment effects for upstream and downstream elections by experiment ........253
List of Tables

Table 2.1: Description of Study 1 treatment arms.................................................................20
Table 2.2: Description of Study 2 treatment arms.................................................................22
Table 2.3: Initial model specifications (50,000 iterations)..................................................36
Table 2.4: Additional model specifications (15,000 additional iterations).........................37
Table 2.5: Summary statistics ..............................................................................................41
Table 2.6: Study 1 – Participants’ Political Party versus their Guess of the Fictional Politician’s Political Party.................................................................42
Table 2.7: Study 2 - Hand-coded Romney facts across conditions ....................................45
Table 2.8: Study 1 – Information about deepfakes has no impact on favorability, vote choice, or support for essential oils in medicine (but we are underpowered) ...........................................................................46
Table 2.9: Study 1 – Deepfakes have no impact on favorability, vote choice, or support for essential oils in medicine (but we are underpowered) ...........................................................................47
Table 2.10: Study 2 - Alternative measure of trust in video as a source of information ........48
Table 2.11: Study 2 – Republicans appear to have directional (non-significant) increases in favorability when information about deepfakes is added to Romney video ............................................48
Table 2.12: Study 2 – Participants who are pro-life appear to have directional (non-significant) increases in favorability when information about deepfakes is added to Romney video ...............49
Table 2.13: Study 2 – No evidence of motivated reasoning ..................................................50
Table 2.14: Study 2 – Some evidence of more distrust immediately before a major election ....51
Table 2.15: Study 2 – No evidence that the information treatment effect is different before versus after a presidential election ........................................................................................................52
Table 3.1: A description of the six overlapping samples analyzed..........................................121
Table 3.2: Hate bands by country..........................................................................................146
Table 3.3: Top listeners of hate music after exclusions.........................................................147
Table 3.4: Friends of top listeners of hate music after exclusions ........................................148
Table 3.5: Trump effects of planned events ................................................................. 153
Table 3.6: Trump effects of unanticipated events ........................................................ 153
Table 3.7: Trump effects of unanticipated and planned events ..................................... 154
Table 3.8: Anticipation effects for planned events among top listeners of hate music .... 154
Table 3.9: Impact of planned events on non-hate music listenership .............................. 163
Table 3.10: Impact of unanticipated events on non-hate music listenership ................. 163
Table 3.11: Impact of unanticipated and planned events on non-hate music listenership . 164
Table 4.1: Overview of all experiments ........................................................................... 224
Table 4.2: Summary statistics (demographics) ................................................................. 226
Table 4.3: Summary statistics (vote history) ................................................................. 227
Table 4.4: Upstream effects by each experiment ............................................................... 232
Table 4.5: Summary of predictions & results across multiple studies ............................ 235
Table 4.6: CACEs and election type ............................................................................... 237
Table 4.7: Salience and downstream CACEs ................................................................. 239
Table 4.8: Race Outcome and CACEs ............................................................................ 240
Table 4.9: Upstream intent-to-treat effect and CACEs for downstream elections by experiment ... 254
This dissertation is dedicated to illuminating the memory of my grandmother, Praskovia Kotlyarova, who lived through a Nazi invasion, Stalinism, and all the consequences of authoritarianism. Her humble heroism inspired me to strive for a better future.
Acknowledgements

I am grateful to everyone who has given me support, advice, and mentorship along the way. A full accounting of acknowledgements would be as lengthy as my dissertation, so I will be succinct.

I thank the chair of my dissertation committee, Peter Aronow, whose unyielding dedication to finding truth in the often-messy world of social science has always been an inspiration. I am grateful to Josh Kalla whose industriousness and diligence has led to critical refinements in my work and research. I thank Greg Huber for somehow making time amid seemingly hundreds of commitments to provide astoundingly succinct yet invaluable advice on where to take my research. I thank Dave Rand, who provided an interdisciplinary perspective on my work, and whose enthusiasm and creativity helped motivate many of my research questions. I am also deeply thankful to Todd Rogers, who provided me with over a decade of academic and professional mentorship.

I thank Taha Yasseri for all the advice and support during my master’s program at Oxford, which helped set the stage for the research in this dissertation. I thank Alex Coppock for the many helpful discussions on online survey experiments and causal inference. I also thank all of my teachers who helped further my methodological and substantive foundations to undertake these projects, particularly Forrest Crawford, Fredrik Sävje, Dan Mattingly, Milan Svolik, Alex Debs, Winston Lin, and Volodymyr Kulyk. I would also like to thank Jay Emerson for all his support in allowing me to earn a master’s in Statistics en route. I’d also like to acknowledge all the helpful feedback I received at the workshops and conferences where this research was presented. I thank Lani Colianna and Colleen Amaro for helping me surmount so many administrative hurdles with their responsiveness, attentiveness, and expertise. I am also grateful to the undergraduate research assistance of Grace Kang and Rosa Kleinman.
I thank my friends and colleagues in brainstorming and thinking through many substantive and methodological questions, particularly Chris Kennedy, Sam Skowronek, Monica Lee, Gonzalo Pons, Justin Kempley, Kyle Peyton, Lilla Orr, Jessica Lasky-Fink, and Georgi Kantchev.

Finally, I am thankful for all the support from my father (whose intellectual rigor embodies everything a scientist should aspire to), my mother (whose self-sacrifice for her family is an inspiration), and my life-partner, Lori Jones (whose intellectual and emotional support has been constant and unflagging).
1. Introduction

Technological advances have begun to change critical aspects of the American political system. The increased connectedness of our systems of communication has led to a world where political misinformation disseminates more rapidly than ever (Vosoughi, Roy & Aral 2017), while Americans’ trust in the traditional news media is at historic lows (Ladd 2011). Some scholars claim that we have now entered an era of “post-truth” or “post-fact” politics (e.g., Mihailidis, & Viotty, 2017; Higgins, 2016). Coupled with the recent resurgence of far-right ideologies, these developments have led some scholars to claim that America is in danger of backsliding out of democracy (Mickey, Levitsky & Way, 2017). The unprecedented storming of the US Capitol building in January 2021 by right-wing rioters highlight the reality of that danger. These circumstances leave a critical need for political science to rigorously address key questions relating to American political sentiments and behavior: How does the rapidly changing media environment affect how Americans receive and trust political information? Do far-right groups in America make use of highly publicized events to recruit others? What motivates Americans to stay politically engaged and continue to turn out to vote? Answering these research questions can both provide invaluable guidance to public policy and allow us to make better sense of longstanding puzzles in political science.

Though the rapidly changing political environment has resulted in difficult political challenges, the silver lining is that new technological developments have allowed political scientists to make use of new disciplines and data sources to produce research that would have been all but impossible just two decades ago. Many of these opportunities require the
tools and knowledge of growing interdisciplinary fields, including data science, computational social science, and judgement-and-decision-making science. As part of this dissertation, I provide three research studies that illustrate the importance of leveraging advances in interdisciplinary fields to answer pressing political science questions. In this chapter, I provide an overview of each of my three papers in turn. For each paper, I present the motivating research question, explain why it is of timely importance, and illustrate how interdisciplinarity opens unique opportunities for political science research on that question.

1.1: Does the Rise of Deepfakes Affect Political Accountability?

Massive increases in partisan polarization (e.g., Layman & Carsey, 2002) and high levels of uninformed voters (e.g., Bartels, 1996) have long been viewed with some alarm (e.g., Achen & Bartels, 2016; Hacker & Pierson, 2019), but empirical research has generally found that the influence of irrational voters tends to wash out in aggregate (e.g., Healy & Malhotra, 2013) and even a polarized electorate will ultimately prefer politicians who reflect their policy preferences (e.g., Costa, 2021). In other words, even in the presence of these challenges, the policy choices made by politicians generally reflect the policy preferences of the electorate. However, democracies can only overcome these issues if key institutional fundamentals are in place. For one, voters need to receive and trust information about what politicians do in office to be able to hold politicians accountable for their actions. If voters, on aggregate, are unable to learn the policy positions of a politician, political accountability breaks down (e.g., Warren, 2014).

A new development in machine learning now allows tech-savvy users to create highly-convincing videos of public figures saying something they’ve never said. These videos are commonly known as “deepfakes” and they have been the cause of alarm for policy-
makers and scholars alike (Dack 2019; O’Sullivan 2019). Unlike textual misinformation, where any attentive individual can pick up clear indications that something is “fake news” (Pennycook et al. 2021), deepfakes have become nearly impossible to detect with the naked eye (see Chapter 2). And though malicious political deepfakes have, so far, been rare, there has been extensive news coverage of the threat of deepfakes (e.g., Gosse & Burkell 2020; Yadlin-Segal & Oppenheim 2021). If this news coverage begins to make the typical American skeptical of direct video footage of politicians making policy statements, and since Americans are already distrustful of the news media (Ladd 2011), the pathways by which real political information reaches the electorate are substantially limited. Political deepfakes and news coverage of deepfakes is not currently widespread, but due to the magnitude of the threat to foundational mechanisms underlying a functioning democracy, there is a clear need to study the impacts of deepfakes on political accountability—especially while there is a substantial “control” group of individuals still unfamiliar with deepfakes.

To adequately address this research question, political scientists must borrow techniques from data science. For one, researchers who are interested in testing the impact of a deepfake that doesn’t already exist in the wider environment must use the most up to date machine learning algorithms to manufacture a deepfake. But in doing so, they should also borrow and develop ethical frameworks from the data science and computer science communities (e.g., Thomas et al. 2017). Since there are still few peer-reviewed studies on the social impact of deepfakes (Dobber at al. 2020; Vaccari & Chadwick 2020), it is critical to balance the needs of research transparency with potential social harms. Namely, deepfake researchers must be wary of unintentionally introducing convincing misinformation into the broader environment. Our colleagues in data science and computer science have much greater experience with similar issues (e.g., Poor and Davidson, 2017), and it would be of
mutual benefit to make use of that knowledge to inform research design decisions to ensure that the social impacts of new technologies are studied ethically.

In Chapter 2 of my thesis, I present a paper that looks at direct and indirect impacts of deepfakes by both applying machine learning algorithms to create a deepfake and making use of ethical principles from data science to ensure that my study clearly avoids ethical gray areas.

1.2: Do Far-Right Hate Groups Use Newsworthy Trump-Related Events to Recruit Others?

Former President Donald Trump’s populist speeches and tweets have been linked to a resurgence of far-right and neo-Nazi movements in America even in the early years of his presidency (e.g. see Bass, 2018; Raghunathan, 2018). The consequences of these alleged links were poignantly felt after the storming of the Capitol in January 2021, as many of the rioters included “self-described Nazis and white supremacists” (Diaz & Treisman, 2021). While the former president was successfully impeached for the “incitement of insurrection” that resulted in the Capitol riot (Naylor, 2021), it is still unclear whether xenophobic, sexist, and racist rhetoric from a President (or a Republican nominee) actually increased far-right group membership. After all, one alternative explanation is that the same complex social processes that led to Trump’s electoral victory in 2016 also led to the growth in far-right hate groups.

While this question has some immediate relevance to, for instance, Trump’s recent de-platforming from Twitter, answering this question can lead to more generalizable knowledge. Particularly, to what an extent are prominent public figures, who themselves are not members of far-right hate groups, nevertheless used as springboards for the recruitment efforts of far-right groups? Research on this question can increase our understanding of how
far-right groups grow and operate, while potentially leading to actionable policy insights as to how their growth and influence can be inhibited.

The key issue is that it is extremely difficult to empirically link a statement Trump made to hate group membership. Many far-right organizations implicitly or explicitly endorse racial violence, which means that membership is not widely publicized. Even organizations dedicated to tracking hate groups in America have not been able to estimate hate group membership precisely enough to measure changes in membership over narrow windows of time (e.g., Southern Poverty Law Center 2020). However, new technological developments have led to the proliferation of large-N behavioral trace data on social networks. In the field of computational social science, there is a growing number of tools and approaches to collecting and making sense of the data that is available. The use of these new datasets has led to innovative research on hate groups and hate crime. For instance, through an instrumental variable design, Muller and Schwarz (2018) estimated the effect of Twitter use on hate crimes, while other researchers made use of far-right online forum posts as a time-series dataset (Scrivens et al. 2020). Furthermore, by consolidating qualitative research with computational social science, it is possible to make use of novel datasets to gain entirely new perspectives on this question.

Specifically, in Chapter 3, I draw on anthropological and sociological research, which found that far-right groups make extensive use of hate music (e.g., bands with such names as “Aryan Terrorism”) for recruitment, organizing, and expression of racist beliefs. I consolidated this qualitative research with techniques in computational social science. Specifically, I made use of Last.fm, one of the largest public repositories for music listenership in the world, to scrape its API (Application Programming Interface), the Last.fm
website, and a historic version of Last.fm on archive.org to generate a dataset of over half a billion song plays from ~250,000 Last.fm users. Such a large-scale collection of data would simply be extremely time-intensive without innovations in computational social science, which has allowed for researchers to collect online data at scale. I also make use of developments in network science in analyzing recruitment processes of far-right groups by, for instance, looking at network centrality measures of users in a far-right music listeners' social network.

1.3: What Motivates Americans to Continue to Turn Out to Vote?

Former president Donald Trump’s attempt to discredit the legitimacy of the 2020 election may have long-term consequences for future Republican voter turnout. Indeed, as-yet unpublished research seems to indicate that the residents of the state of Georgia who most strongly believe in the “election fraud” conspiracy on Twitter were less likely to vote in the 2021 Georgia runoff election (Green et al. 2021). Are attempts to discredit the electoral process likely to be self-defeating or are the consequences for voter turnout more complex? This question highlights the importance of understanding the mechanism behind why an individual turns out to vote in the first place.

Though the political behavior literature on voter turnout is extensive (see Green & Gerber 2019 for an overview), there is one piece of empirical evidence on voting behavior that is not yet integrated within a cohesive theoretic framework. Namely, there is strong, consistent evidence showing that turning out in one election makes the same individual more likely to turn out in a subsequent election (e.g., Coppock & Green 2016). Some scholars have concluded that this phenomenon implies that voting is “habit-forming,” but turnout persistence fails to comport with habit as it is defined by psychologists—habit is an
automatic behavior that forms from frequent repetition of an activity (e.g., Dinas 2012).
Fortunately, the growing field of judgement and decision-making (JDM) science has
developed important research that filled the gaps between social psychology and more
applied social science fields such as economics and political science (Duckworth & Gross
2020). By leveraging developments in this interdisciplinary cognitive science, it is possible to
bridge a gap in the literature between conventional economic models of voting and the
voting-as-a-habit literature. In Chapter 4, I illustrate the usefulness of applying theories
developed in the JDM literature by applying a new JDM model (Duckworth and Gross’s
(2020) Process Model for Behavior Change) to an existing puzzle in political science. I also
make use of the fact that American campaigns and non-profits have increasingly begun to
incorporate field experiments in their Get Out the Vote (GOTV) programs. Namely, I
partnered with a large labor organization to test the predictions of the JDM model as applied
to turnout persistence across 9 independent experiments with a total of ~1.8 million voters.

1.4: References

2018. https://www.npr.org/2018/01/20/579443004/member-of-congressional-
black-caucus-trump-has-brought-normalized-racism-to-whit
experiments and regression discontinuities. American Journal of Political Science, 60(4),
1044-1062.


2. The Negative Consequences of Informing Voters about Deepfakes: Evidence from Two Survey Experiments

2.1: Significance Statement

We evaluate the social impacts of a new development in machine learning: computer-manipulated video misinformation (so-called “deepfakes”). While prior research focused on political deepfakes’ limited capacity to mislead, this research, in contrast, explores important second-order effects: what happens when voters are informed about the existence of seemingly undetectable manipulated videos? Is the public’s confidence in video as a source of political information undermined? These questions have serious consequences for the health of a democracy. Many Americans distrust the news media; if they begin to disbelieve real videos of politicians, they are left with few trusted sources for factual information about politicians’ statements and actions in office, which jeopardizes their ability to vote out bad actors.

2.2: Abstract

Advances in machine learning have made possible “deepfakes,” or realistic, computer-generated videos of public figures saying something they have not actually said. Policymakers

* This paper was co-authored with Joshua Kalla and P.M. Aronow. John Ternovski was responsible for all computing, data collection and analysis, and wrote the initial manuscript and supplementary materials. All authors contributed to conceptualization, study design, and revision of the manuscript. This research was approved by the Yale University IRB. We thank Gregory A. Huber, David G. Rand, and Todd Rogers for useful discussions. This paper was presented at the Harvard Experimental Political Science Conference, MIT Sloan, Yale University, and Dartmouth College; we thank all involved for the opportunity to present and helpful feedback. We are grateful to Grace Kang and Rosa Kleinman for their research assistance.
have expressed concern that deepfakes could mislead voters, but existing research has found minimal effects. There has nevertheless been extensive media coverage on the dangers of deepfakes, urging voters to be critical consumers of political video. We explore whether these well-intentioned activities have an unintended consequence: if voters are warned about deepfakes, they may begin to distrust all political video. Through two online survey experiments, we found that informing participants about deepfakes did not enhance participants’ ability to successfully spot manipulated videos but consistently induced participants to believe that the videos they watched were fake, even when the videos were real. Our findings suggest that even if deepfakes are not themselves persuasive, information about the existence of deepfakes can nevertheless be weaponized to dismiss real political video.

2.3: Introduction

Breakthroughs in machine learning have led to the development of software that can seamlessly fabricate videos of any given individual. Computer-generated videos, so-called “deepfakes,” can be made in which a politician appears to say something they have never actually said in real life.¹ Computer and social scientists have raised concerns that deepfakes may mislead voters and sway election outcomes (e.g., Dack 2019). Policymakers have echoed these concerns. For example, during a hearing of the United States House of Representatives Intelligence Committee, Adam Schiff, the Committee’s chair, noted that deepfakes allow “malicious actors to foment chaos, division or crisis,” and that such videos “have the capacity to disrupt entire campaigns, including that for the presidency” (O’Sullivan 2019, 1). Since the hearing, Congress has passed two laws (IOGAN Act 2020; NDAA for FY2021

¹ For a sociological overview of the development of deepfakes, see Paris & Donovan (2019).
that explicitly directed “the Department of Homeland Security (DHS), the Department of Defense (DOD), and the National Science Foundation (NSF) to issue reports on and bolster research into deepfakes… These bills ask for recommendations that could lay the predicate for federal regulations of such media.” (Ferraro, 2020, 1) However, recent randomized experiments on the impacts of deepfakes in American politics have found no evidence that people believe the content of the manipulated videos (Wittenberg, Zong, & Rand 2020; Vaccari & Chadwick 2020). More recently, Barari, Lucas & Munger (2021) did find that deepfakes were persuasive but their effects were comparable to that of textual misinformation.2

Despite this evidence from social scientists, news coverage of deepfakes continues to be extensive and predominantly emphasizes the threat of deepfakes (e.g., Gosse & Burkell 2020; Yadlin-Segal & Oppenheim 2021). A cursory search of the five most popular news websites in the US (according to YouGov (2021)) for the search term “deepfake”3 on the news aggregator Google News finds 11,700 news articles discussing deepfakes and 62.5% of those articles use cautionary language (“threat”, “worried”, “danger”, “warn”, “risk”).4,5 Attempts to inform the public of the dangers of deepfakes even led to the creation of a widely-viewed deepfake of former President Barack Obama, in which the comedian Jordan

---

2 There is also a recent deepfake study in the Netherlands that found modest persuasive effects among a subset of participants (Dobber at al. 2020).

3 The search was conducted on March 30th, 2021 with the input: deepfake AND (site:news.yahoo.com OR sitecnbc.com OR sitecbs.com OR sitecnbcnews.com OR sitecnnc.com)

4 The search was conducted on March 30th, 2021 with the input: deepfake AND (“threat*” OR "worried" OR "danger*" OR "warn*" OR "risk*") AND (site:news.yahoo.com OR sitecnbc.com OR sitecbs.com OR sitecnbcnews.com OR sitecnnc.com)

5 We qualitatively assessed these search results. While there were, unsurprisingly, many false positives for both search queries, we did find that, in concordance with Gosse & Burkell (2020) and Yadlin-Segal & Oppenheim (2021), the overwhelming majority of articles that did specifically address deepfakes focused on their potential danger or, at best, described them as “creepy.”
Peele partnered with BuzzFeed Video to create a deepfake of himself impersonating Obama to warn Americans about deepfakes (Castillo 2018). This public service announcement has since accrued over 8.4 million views on YouTube. But do these well-intentioned attempts to inform and educate the public have an unintended consequence: does information about the existence and dangers of deepfakes cause voters to distrust all political video footage—whether real or fake?

This question has not been answered in the context of deepfakes, but a robust literature on textual misinformation (for a review, see Lazer et al. 2018) has found that elite discourse about “fake news” may lower trust in the media and prime participants to disbelieve the veracity of real news (Van Duyn & Collier 2019). Official warnings about fake news similarly induce participants to disbelieve true headlines (Clayton et al. 2020; Pennycook, Bear et al. 2020). There are already high-profile cases of American voters alleging that real political videos are deepfakes. For example, in January 2021, supporters of Donald Trump suggested that a video Trump shared via Twitter in which he conceded the 2020 election was a deepfake (Villarreal 2021).

It isn’t only highly-motivated partisans that disbelieve real video footage. Recently, there was a high-profile case of the American justice system erroneously alleging that real video footage was deepfaked. Namely, a Pennsylvania woman was accused of making a deepfake of high school cheerleaders vaping “to try to get them kicked off the squad” (Associated Press, 2021), but upon a closer examination, video forensic experts found no evidence that the video was manipulated (Harwell 2021). “When pressed on how police made their determination that the footage had been manipulated… [the police officer who made the arrest said] that he had relied on his ‘naked eye.” (Thalen 2021). Without the law
enforcement officer’s awareness of the existence of deepfake technology, such an accusation would have been all but impossible.

This paper explores whether information warning about the existence of deepfakes makes American voters more likely to disbelieve real political videos. Or do these efforts to inform voters about the dangers of deepfakes work as intended—leading to a more critical consumption of political video? If the former is true, as media coverage of deepfakes continues to increase, Americans’ trust in political video may continue to erode. The other danger is that politicians could use factually true statements (e.g., “deepfakes exist”) to subtly disavow and dismiss video recordings of their past statements and behavior. Such outcomes could potentially undermine political accountability and so it is imperative to understand the social impacts of information about deepfakes before political deepfakes are commonplace.

Across two online survey experiments, we demonstrate that informing voters about deepfakes increases disbelief in both real and manipulated videos without improving participants’ ability to successfully identify the deepfake. Study 1 used an actor posing as a politician sharing an extreme policy position. Using a factorial design, participants were randomized in a first factor to either see a real video of the politician or a deepfake version of the video and, in a second factor, to receive information about deepfakes or not. Study 2 made use of Americans’ low levels of policy knowledge (e.g., Barabas et al. 2014) to show a real video of a real-world politician making a policy statement that is atypical of his party and is not widely known (and thus might reasonably be inferred by voters to be a deepfake). Both studies measured belief in the content of the videos and trust in video as a source of political information.

We found that participants were unable to discriminate between real and deepfaked video even when they were informed about the existence of deepfakes. Instead, information
about deepfakes induced participants to disbelieve any associated political video—real or fabricated. In other words, a general statement about the dangers of deepfakes, as one might see in a headline from a trusted news source, was enough to nudge participants to disbelieve real video clips of politicians making policy statements. The effects were large and consistent across both studies. Information about deepfakes even affected what policy stances participants associated with a real-world US politician. In other words, providing information about the dangers of deepfakes not only made participants suspect that real videos of politicians speaking are fake, but even affected how the content of the video is internalized.

This paper is organized as follows. First, we present a brief overview of the data and the designs of both survey experiments. Next, we report our main findings and discuss the implications of the results. We conclude with broader policy implications and delineate avenues for further research.

2.4: Experimental Design

We ran two pre-registered6 online survey experiments on Lucid Theorem7 using convenience samples8 in the spring and fall of 2020. The full survey questionnaires are provided in the Supplementary Materials.

---

6 See https://osf.io/rqfr5/?view_only=e2807b367a534262bb6ce7aeb5727b999 for our pre-analysis plans.

7 Lucid is an increasingly popular alternative to Amazon Mechanical Turk for social science survey research. Many well-known findings have been replicated on Lucid, suggesting the platform is capable of providing high-quality data (Coppock & McClellan 2019, Peyton, Huber, & Coppock 2020). During the COVID-19 pandemic, Aronow et al. (2020) found that Lucid data can provide reliable data when researchers screen on attentiveness, which we do here.

8 Treatment effects from online convenience samples have been shown to generalize to nationally representative samples (e.g., Mullinix et al., 2015).
2.4.1: Study 1 - Fictional Politician

We created a video of an actor playing a fictional politician advocating an extreme policy position: support for a law requiring doctors to use essential oils to treat cancer before attempting treatments with conventional medicine.\(^9\) (A screenshot of the video can be viewed in Figure 2.1.) The actor also recorded another video that we used as training footage for our deepfake, where he recited a generic political speech.\(^{10}\) We used this visual data to construct a SAEHD (High Definition Styled AutoEncoder) model (trained over 65,000 iterations). The resulting model is essentially a moldable mask that was superimposed on the actor in the “destination” video (i.e., the video where our hired actor talks about essential oils in medicine). The full technical details of the deepfake model are found in the Supplementary Materials (2.8.1). This process emulates a malicious agent using a look-alike actor as a training set for the deepfake. The real video on essential oils and the deepfake thereof were used as treatments in Study 1.

---

\(^9\) We confirmed that the fictional politician appeared sufficiently non-partisan by asking participants to guess the politician’s political party at the end of the survey. The plurality of respondents (30%) said they “didn’t know” (or it was “unclear” as to) what political party the politician was affiliated with. 27% thought the politician was a Democrat, 16% thought the politician was a Republican, and 27% thought the politician was Independent. The partisanship of respondents did interact with respondents’ guesses, but the relationship was weak (Cramer’s $V = 0.08$; see Table 2.6 in the Supplementary Materials for more details).

\(^{10}\) The speech was composed using excerpts from Swaim & Nussbaum (2016), a FiveThirtyEight article wherein Republican and Democratic speechwriters collaborated “to write a totally pandering bipartisan stump speech for an imaginary presidential candidate — one who espouses only positions that a majority of voters agree with” (p. 1).
Study 1 began with a series of screener and pre-treatment measures (for details, please see the survey script in the Supplementary Materials). Participants were then shown a separate page with the following text: “On the next page, you will watch a brief political speech. You will then be asked about how you feel about the person making the speech and whether or not you would vote for them in the upcoming election.” We randomized participants to receive either the real video or the deepfake version and either information about deepfakes or no additional information in a 2x2 fully factorialized experiment. The four conditions are summarized in Table 2.1.
Table 2.1: Description of Study 1 treatment arms

<table>
<thead>
<tr>
<th>Real Video (N=696)</th>
<th>No Information (N=696)</th>
<th>Information about Deepfakes (N=700)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>Participants were shown a real video of the actor stating that, if elected, he would mandate doctors to try using essential oils to treat cancer before they are allowed to use conventional medicine. (N=347)</td>
<td>Participants were shown the same video as (1) but with the following message above the clip “WARNING: Computer scientists are increasingly concerned about &quot;DeepFake&quot; videos. With DeepFakes, it's now possible to use a computer to convincingly manipulate videos of people to make them appear to say things they've never said.” (N=349)</td>
</tr>
<tr>
<td>Deepfake Video (N=700)</td>
<td>(3) This treatment arm was identical to (1) except participants were shown the deepfake version of the video. (N=349)</td>
<td>(4) This treatment arm was identical to (2) except participants were shown the deepfake version of the video. (N=351)</td>
</tr>
</tbody>
</table>

After viewing the video, we measured participants’ perceived favorability of and intention to vote for the politician, their view on essential oils in medicine, whether they believed the video was real, and their overall confidence in other video footage of politicians speaking. Due to the risk of priming latent beliefs by asking participants if the video they just saw was real,11 we attempted to measure this outcome unobtrusively by asking whether they were convinced that the politician “believes what is said.” We only explicitly ask participants whether they believed the video was deepfaked at the very end of the survey.

---

11 Research on textual fake news found that asking participants about the accuracy of a specific piece of misinformation affected how likely they were to share that misinformation online (Pennycook et al. 2021; Pennycook, McPhetres et al. 2020). If priming accuracy can affect behavior, it has the potential to affect latent beliefs.
In this study, we found evidence that survey attrition varied significantly across treatment conditions (Pearson chi²(3) = 11.3, p = 0.01). 4.9% participants in the No Information + Real Video condition did not finish the survey compared to 1.8% in the aggregate sum of the three remaining conditions. Because our pre-analysis plan did not explicitly stipulate how we would address differential attrition, we used Manski-type worst-case bounds (Manski 2003) that require only that the support of the outcome is bounded when constructing our confidence intervals. These bounds are considered the gold-standard approach when analyzing experiments with attrition (Gerber and Green, 2012). Additional details are provided in Section 2 of the Supplementary Materials.

2.4.2: Study 2 - Real Politician

One weakness of Study 1 is that the extreme policy stance in the video could plausibly be genuine. Participants were not familiar with the politician (as he is fictional) and may not have had any expectation that the video they were watching could be deepfaked. McDonald (2019) empirically illustrated how people make use of prior knowledge of real-world politicians in online survey experiments and how studies using solely hypothetical politicians can produce misleading estimates of real-world political behavior. We therefore replicated the impact of providing information about deepfakes using a real politician in Study 2.

We identified a well-known politician who, at one point in his political career, had expressed a policy stance that was atypical given his party affiliation. Specifically, we found 2002 video footage of Republican Mitt Romney asserting, in a Massachusetts gubernatorial

---

12 The “best” and “worst” cases used to estimate the CI are also covariate adjusted and use robust standard errors.
debate, that he would protect a woman’s right to choose – an unusual stance for a Republican.¹³

The design of this study was nearly identical to Study 1 with the following modifications. First, Study 2 had only two conditions, Information about Deepfakes and No Information (see Table 2.2 for a description of the two conditions), as we did not create a deepfake of Mitt Romney given ethical considerations. Second, our analysis was restricted to participants who knew Romney was a Republican pre-treatment, as Romney’s (formerly) pro-choice position would not appear surprising to those participants with no knowledge of Romney’s party affiliation. In Study 2, we found no evidence of differential attrition: survey completion rates did not differ significantly across treatment arms.

Table 2.2: Description of Study 2 treatment arms

<table>
<thead>
<tr>
<th>Real Video (N=1,925)</th>
<th>No Information (N=966)</th>
<th>Information about Deepfakes (N=959)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Participants watched a real video of Mitt Romney saying, “And I’ve been very clear on that, I will preserve and protect a woman’s right to choose and I’m devoted and dedicated to honoring my word…”</td>
<td>(2) Participants were shown the same video as (1) but with the following message above the clip “WARNING: Computer scientists are increasingly concerned about &quot;DeepFake&quot; videos. With DeepFakes, it’s now possible to use a computer to convincingly manipulate videos of people to make them appear to say things they’ve never said.”</td>
<td></td>
</tr>
</tbody>
</table>

¹³ As of 2020, there were only 2 pro-choice Republicans in the Senate (Sussman 2020).
2.5: Results and Discussion

News articles about the dangers of deepfakes and PSAs like Jordan Peele’s Obama deepfake are intended to make viewers more critical when consuming political video. Ideally, a viewer will believe real videos and disregard fake videos. It is possible that the intent of such messages is to make Americans more skeptical of all information; but such a goal could have deleterious impacts on democratic functioning. Belief in fake videos may lead to misinformed voters, but disbelief in real videos of politicians discussing their policy positions may lead to uninformed voters (for further analysis of uninformed and misinformed voters, see Kuklinski et al. 2000). High levels of uninformed voters have been linked to serious electoral consequences; for instance, Fowler and Margolis (2014) found that “[a] lack of knowledge on the policy positions of the parties significantly hinders the ability of low-socioeconomic-status citizens to translate their preferences into partisan opinions and vote choices.” (p. 100).

As such, our first analysis assesses whether informing participants about the danger of deepfakes affects the rate at which voters disbelieve a deepfaked political video. The primary outcome measure used asked participants how much they agreed with the statement “This video was doctored, manipulated and/or faked by a computer (i.e. it is a ‘Deep Fake’)” on a 7-point agree/disagree scale ranging from strongly disagree (-3) to strongly agree (3).14 In the leftmost column of Figure 2.2, we see that when the information about deepfakes was randomly added to a deepfake video, participants were 0.5 points more likely to believe that

---

14 Alternative survey instruments found similar (but smaller) treatment effects for all analyses. The results of these alternative outcome measures can be viewed in Figures 2.4 and 2.5 in the Supplementary Materials.
the video they were watching was fabricated (p<0.001).\textsuperscript{15} It is worth noting that this treatment effect was not driven by people affirmatively identifying the deepfake. When participants watched a deepfaked video without information about deepfakes, they were fairly confident that what they were watching was not a deepfake (-0.4 points on our 7-point scale or approximately halfway between “somewhat disagree that the video is a deepfake” and “neither agree nor disagree that the video is a deepfake”); when information about deepfakes was added, they became more uncertain about whether the video was a deepfake (0.1 points).\textsuperscript{16} This result is consistent with Vaccari & Chadwick’s (2020) conclusions that deepfakes increase uncertainty.

\textsuperscript{15} Cohen’s d of this effect is 0.3, which suggests that this effect size is somewhere between small and medium (Cohen 2013).

\textsuperscript{16} Participants were not able to discern between a real video and a deepfake version of the same video without the Information treatment (see Section 3 of the Supplementary Materials for more details).
Figure 2.2: Information about deepfakes induced disbelief in accompanying video regardless of whether the video is real or fake.

The top row illustrates covariate unadjusted mean sentiment in each treatment arm, while the bottom row shows the resulting treatment effects with 95% confidence intervals. The leftmost column illustrates the impact of information about deepfakes added above a deepfake video; the remaining columns illustrate the impact of information about deepfakes added above a real video. All three columns measure impact in terms of the same outcome: do you believe the video clip is a deepfake? [7-point scale from -3="strongly disagree" to 3="strongly agree"]). The two leftmost treatment effect graphs use OLS estimates of Manski-type worst-case bounds with robust standard errors and covariate adjustment to calculate 95% confidence intervals. The rightmost treatment effect graph uses OLS estimates with robust standard errors and covariate adjustment, as there was no differential attrition.
The next step is to investigate whether the effect of providing information about deepfakes has the unintended consequence of reducing belief in real video. We see a comparable treatment effect (.75 points [p<0.001]) when information about deepfakes was randomly added to a real video of the same content (see the middle column in Figure 2.2).\textsuperscript{17} The interaction of the deepfake treatment and the information treatment is not statistically significant.\textsuperscript{18} Rather than helping participants detect deepfake videos, information about deepfakes instead caused participants to disbelieve whatever video they were watching—real or fake. This interaction effect is directionally opposite from the normative ideal: adding information about deepfakes to videos made participants more likely to disbelieve the real video rather than successfully identify the deepfake.

We find similar effects with the real video of Romney in Study 2. The results are summarized in the rightmost column of Figure 2.2. As before, we see that information about deepfakes induced participants to disbelieve real videos—an effect of 0.26 points on a 7-point scale (p<0.001).\textsuperscript{19}

Pooling the data from the two studies and including a study fixed effect, we find that information about deepfakes increased participants’ belief that the video was fabricated by 0.40 points on a 7-point scale (p<0.001).\textsuperscript{20} Information about deepfakes also increased how unconvinced participants were that the politician actually believed what he was saying by

\textsuperscript{17} Cohen’s d of this effect is 0.5, which suggests that this is a medium effect size (Cohen 2013).

\textsuperscript{18} Even without bounds, the p-value is 0.13 and the effect is in the opposite direction from what would be desired.

\textsuperscript{19} Cohen’s d of this effect is 0.2, which suggests that this is a small (but nontrivial) effect size (Cohen 2013).

\textsuperscript{20} We exclude Study 1 participants randomized to the deepfake condition in the reported specification, but the results do not change meaningfully when we include the deepfake condition.
0.13 points on a 3-point scale (p<0.001). Finally, information about deepfakes also increased the rate at which participants said that they “didn’t know whether the politician believed what was said in the video” (a binary variable) by 6.3 percentage points (p=0.001).

We also evaluated whether information about deepfakes can affect how participants internalize the content of the video. Namely, does information about deepfakes change what facts participants associate with the politician in the video? Towards the end of Study 2, we asked participants to name three facts about Romney in an open-ended question. We found that information about deepfakes did not significantly change how likely participants were to mention abortion (p=0.68), but there was a major shift in what participants perceived Romney’s position on abortion to be. The information about deepfakes caused a 3.0 percentage point drop in the percentage of participants who associated Romney with a pro-choice position (p=0.001). (For more details on the open-ended question, please see Table 2.7 in the Supplementary Materials.) An alternative measure of this outcome asked participants in a close-ended question if Romney had ever “supported women’s access to abortion”; we found that adding information about deepfakes decreased the likelihood that participants marked “true” by 14.2 percentage points (p<0.001).

2.5.1: Heterogenous Treatment Effects

We also examined heterogeneous treatment effects, but as we noted in both pre-analysis plans, we were underpowered for most of these analyses. We give a brief overview of the most noteworthy results here (see Sections 4-5 of the Supplementary Materials for full results). First, we found some evidence that participants surveyed before the election have higher levels of distrust in political videos than participants surveyed after the election (p<.05). While we originally planned to investigate motivated reasoning in Study 2, by the
time the study was launched, Romney had already become a polarizing figure among Republicans, so our data does not allow us to adequately address this question. A more thorough discussion of this point can be found in Section 6 of the Supplementary Materials.

### 2.6: Discussion and Conclusions

Our findings add to the growing body of literature on textual misinformation, which has consistently found that warnings about fake news may help readers reject misinformation but may have the undesirable backlash of increasing Americans’ disbelief in true news stories (e.g., Pennycook, Bear et al. 2020; Clayton et al. 2020). We find strong evidence that cautionary information about deepfakes, as one might see in a news headline, increases disbelief in accompanying video clips—regardless of whether the video is fake or real. This is particularly problematic as the information about deepfakes did not state that the accompanying video was fake; the information treatment simply stated that deepfakes exist and are challenging to spot, which is something that an American might hear on a news program. Such a nudge nevertheless induced people to disbelieve a video that revealed a little-known but real policy stance of a real-world politician. And not only did participants suspect that the video was fake, participants’ beliefs about the politician’s policy stances changed as a result of the information treatment.

There are limitations to these findings. The data is sourced from online survey experiments with convenience samples of participants, which means that there are legitimate concerns over the external validity of our results (see Coppock (2019) for a thorough analysis of these general concerns.) We acknowledge that the deepfake information treatment does not resemble how the average person is likely to be exposed to information about deepfakes in the real world. But one advantage of our treatment is that we remove
many of the contextual cues of real world exposure (e.g., news source), which helps isolate
the effect of being informed about deepfakes from other related effects (e.g., having an
emotional reaction to the source of the information). One other notable limitation is that the
time between exposure to information about deepfakes and videos of politicians making
policy statements may be much longer in the real world than in our survey experiments.
Future studies should investigate if and how quickly these treatment effects decay. One
countervailing possibility is that, in the real world, people may receive higher dosages of
deepfake information (e.g., through more extensive news coverage) which may cause larger
increases in skepticism. As such, future research should assess the impact of dosage.

Our results illustrate that, while well-intentioned, attempts to warn the public about
deepfakes may inadvertently cause the delegitimization of true information. Our findings
suggest that the news media, elites, and social media platforms may need to take great care in
their attempts to educate the public. We show that providing information about the
existence and the potential dangers of deepfakes erode trust, and it is thus imperative that
other approaches be the subject of future research.

2.7: References

Aronow, Peter Michael, Joshua Kalla, Lilla Orr, and John Ternovski. 2020. "Evidence of

https://osf.io/8sbe4/

cheerleading-0953a60ab3e3452b87753e81e0e77d7f


Coppock, Alexander, and Oliver A. McClellan. 2019. Validating the demographic, political, psychological, and experimental results obtained from a new source of online survey respondents. *Research & Politics* 6, no. 1: 2053168018822174.

Dobber, Tom, Nadia Metoui, Damian Trilling, Natali Helberger, and Claes de Vreese. 2020. "Do (microtargeted) deepfakes have real effects on political attitudes?" The International Journal of Press/Politics: 1940161220944364.


[https://datasociety.net/library/deepfakes-and-cheap-fakes/](https://datasociety.net/library/deepfakes-and-cheap-fakes/)


[https://psyarxiv.com/r5yun/](https://psyarxiv.com/r5yun/)


[https://today.yougov.com/ratings/media/popularity/news-websites/all](https://today.yougov.com/ratings/media/popularity/news-websites/all)
2.8: Supplementary Materials

2.8.1: Deepfake Model Specifications

We used Amazon’s graphics-accelerated Elastic Computing (EC2) environment to parse the training footage into a training set of still .jpg images of the actor. These images were used to build a deepfake model using a widely available, open-source deepfake software package, DeepFaceLab. The package uses Google’s deep learning library, TensorFlow.

Only the video component was deepfaked—not the audio. This was a conscious choice since vocal impersonations are much easier and often more convincing than audio deepfakes. For example, one of the most widespread deepfakes in the wild (Jordan Peele as Obama) did not deepfake audio; Jordan Peele merely impersonated Barack Obama’s voice.

The actor’s training video was a generic political speech, which was composed using excerpts from a FiveThirtyEight article wherein Republican and Democratic speechwriters collaborated “to write a totally pandering bipartisan stump speech for an imaginary presidential candidate — one who espouses only positions that a majority of voters agree with” (Swaim & Nussbaum, 2016; p. 1).

The deepfake was created on Amazon’s g3s.xlarge instance, which uses a Tesla M60 GPU. Using the 2/3/2020 NVIDIA build of DeepFaceLab, we first ran SAEHD for 50,000 iterations using the following specifications:
Table 2.3: Initial model specifications (50,000 iterations)

<table>
<thead>
<tr>
<th>Setting</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>flip faces</td>
<td>No</td>
</tr>
<tr>
<td>batch size</td>
<td>8</td>
</tr>
<tr>
<td>resolution</td>
<td>128</td>
</tr>
<tr>
<td>face type</td>
<td>F</td>
</tr>
<tr>
<td>AE architecture</td>
<td>dfhd</td>
</tr>
<tr>
<td>autoencoder</td>
<td>512</td>
</tr>
<tr>
<td>encoder dims</td>
<td>64</td>
</tr>
<tr>
<td>decoder dims</td>
<td>48</td>
</tr>
<tr>
<td>decoder mask dim</td>
<td>16</td>
</tr>
<tr>
<td>learn mask</td>
<td>no</td>
</tr>
<tr>
<td>optimizer on GPU</td>
<td>yes</td>
</tr>
<tr>
<td>learning rate dropout</td>
<td>no</td>
</tr>
<tr>
<td>random warp</td>
<td>yes</td>
</tr>
<tr>
<td>GAN power</td>
<td>0</td>
</tr>
<tr>
<td>face style power</td>
<td>0</td>
</tr>
<tr>
<td>background style power</td>
<td>0</td>
</tr>
<tr>
<td>color transfer</td>
<td>none</td>
</tr>
<tr>
<td>gradient clipping</td>
<td>yes</td>
</tr>
<tr>
<td>enable pretraining</td>
<td>no</td>
</tr>
</tbody>
</table>

The above model resulted in flickering, so we proceeded for 15,000 more iterations using the following settings:
Table 2.4: Additional model specifications (15,000 additional iterations)

<table>
<thead>
<tr>
<th>Feature</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>flip faces</td>
<td>no</td>
</tr>
<tr>
<td>batch size</td>
<td>4</td>
</tr>
<tr>
<td>resolution</td>
<td>128</td>
</tr>
<tr>
<td>face type</td>
<td>f</td>
</tr>
<tr>
<td>AE architecture</td>
<td>dfhd</td>
</tr>
<tr>
<td>autoencoder</td>
<td>512</td>
</tr>
<tr>
<td>encoder dims</td>
<td>64</td>
</tr>
<tr>
<td>decoder dims</td>
<td>48</td>
</tr>
<tr>
<td>decoder mask dim</td>
<td>16</td>
</tr>
<tr>
<td>learn mask</td>
<td>yes</td>
</tr>
<tr>
<td>optimizer on GPU</td>
<td>yes</td>
</tr>
<tr>
<td>learning rate dropout</td>
<td>yes</td>
</tr>
<tr>
<td>random warp</td>
<td>no</td>
</tr>
<tr>
<td>GAN power</td>
<td>0</td>
</tr>
<tr>
<td>face style power</td>
<td>0</td>
</tr>
<tr>
<td>background style power</td>
<td>0</td>
</tr>
<tr>
<td>color transfer</td>
<td>none</td>
</tr>
<tr>
<td>gradient clipping</td>
<td>yes</td>
</tr>
<tr>
<td>enable pretraining</td>
<td>no</td>
</tr>
</tbody>
</table>

To merge the model to the destination video, a seamless histogram match was used with the mask eroded by 30.

2.8.2: Deviations from Pre-Analysis Plan (PAP)

2.8.2.1: Study 1

This study had one major and two minor deviations from the PAP.

The major deviation is the use of Manski bounds for confidence intervals and statistical significance. For convenience, we provide a brief summary of the method here (for more details, please refer to Manski (2003)). We impose the distribution of outcomes that least favors our hypothesis onto all missing outcomes (i.e., the survey instrument maximum for all missing outcomes in the control and the survey instrument minimum for all missing
outcomes in treatment), run a covariate-adjusted OLS regression with robust standard errors over all outcomes, and use the lower bound of the resulting 95% confidence interval as the lower bound of our reported 95% CI. We do the same with a best-case distribution of missing outcomes to estimate the upper bound of our reported 95% CI.

We also had two minor deviations from the PAP. To be consistent with Study 2’s PAP, we 1) dropped the one IP address that was randomized three times (there were no other IP duplicates that watched our video) and 2) used robust standard errors. These modifications have negligible impacts on the reported results.

2.8.2.2: Study 2

There were no deviations from the PAP in Study 2.

2.8.2.3: Deviations from the .do Files Accompanying the PAP

There was a typo in the data cleanup code accompanying the PAP. Political party in both studies was originally miscoded such that “Other - Neither Republican or Democrat” was treated as Republican. There were also minor capitalization and naming errors. These were fixed in the final code, which can be viewed in full in the Appendix. Additional changes include dropping pilot data (as described in the original PAPs), de-duping on IP address in Study 1 (as described above), adding an indicator identifying cases where self-reported age was not within one year of age according to Lucid in Study 2 (as described in Study 2’s updated PAP).

2.8.3: Discernment of Deepfakes Without Deepfake Information

One key question is whether American voters can discern between deepfakes and real videos without being informed about deepfakes. In Figure 2.3, we show that participants
were unable to do so. The leftmost column in Figure 2.3 uses an outcome measure that asked participants how much they agreed with the statement “This video was doctored, manipulated and/or faked by a computer (i.e. it is a ‘Deep Fake’)” on a 7-point agree/disagree scale. Participants in the real video condition had a mean sentiment of -0.48, which is nearly equidistant between “somewhat disagree” and “neither agree nor disagree.” The mean sentiment of participants in the deepfake condition was not significantly different from that value (only 0.07 points higher). As shown in the remaining two columns, treatment effects were even smaller across alternative measures of disbelief. In other words, a well-made deepfake video is unlikely to be detected by the naked eye of a typical American voter. In fact, it appears that Americans are somewhat confident that a given political video is not manipulated by a computer—even when it is.
Figure 2.3: Voters were unable to discriminate between a real video and a deepfake (Study 1)

This figure includes only those participants who did not receive a deepfake information. The top row illustrates covariate unadjusted mean sentiment in each treatment arm for each outcome. The bottom row shows covariate-adjusted treatment effects with Manski bounds used to calculate 95% confidence intervals.

The three columns represent three outcomes (from left to right): 1) do you believe the video clip is a deepfake? [7-point scale from -3="strongly disagree" to 3="strongly agree"]; 2) how convinced are you that the politician believes what is being said? [3-point scale from 1="not at all convinced" to -1="very convinced"]; and 3) I don’t know if the politician truly believes what is said [binary variable].
### Table 2.5: Summary statistics

<table>
<thead>
<tr>
<th></th>
<th>Study 1</th>
<th>Study 2</th>
<th>Pooled</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18-29</td>
<td>24%</td>
<td>15%</td>
<td>19%</td>
</tr>
<tr>
<td>30-49</td>
<td>33%</td>
<td>34%</td>
<td>33%</td>
</tr>
<tr>
<td>50-64</td>
<td>27%</td>
<td>28%</td>
<td>27%</td>
</tr>
<tr>
<td>65+</td>
<td>16%</td>
<td>24%</td>
<td>21%</td>
</tr>
<tr>
<td><strong>Gender</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% Female</td>
<td>51%</td>
<td>54%</td>
<td>53%</td>
</tr>
<tr>
<td><strong>Race</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>73%</td>
<td>79%</td>
<td>77%</td>
</tr>
<tr>
<td>Black</td>
<td>11%</td>
<td>9%</td>
<td>10%</td>
</tr>
<tr>
<td>Hispanic</td>
<td>12%</td>
<td>6%</td>
<td>8%</td>
</tr>
<tr>
<td><strong>Region</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>South</td>
<td>38%</td>
<td>38%</td>
<td>38%</td>
</tr>
<tr>
<td>Midwest</td>
<td>19%</td>
<td>20%</td>
<td>20%</td>
</tr>
<tr>
<td>Northeast</td>
<td>21%</td>
<td>22%</td>
<td>21%</td>
</tr>
<tr>
<td>West</td>
<td>22%</td>
<td>20%</td>
<td>21%</td>
</tr>
<tr>
<td><strong>Education</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HS or less</td>
<td>18%</td>
<td>16%</td>
<td>17%</td>
</tr>
<tr>
<td>Some College</td>
<td>37%</td>
<td>29%</td>
<td>32%</td>
</tr>
<tr>
<td>Bachelor's Degree</td>
<td>27%</td>
<td>29%</td>
<td>29%</td>
</tr>
<tr>
<td>Postgrad</td>
<td>17%</td>
<td>25%</td>
<td>22%</td>
</tr>
<tr>
<td><strong>Partisanship</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Identify or Lean Democrat</td>
<td>48%</td>
<td>48%</td>
<td>48%</td>
</tr>
<tr>
<td>Identify or Lean Republican</td>
<td>39%</td>
<td>43%</td>
<td>41%</td>
</tr>
<tr>
<td><strong>Median Household Income</strong></td>
<td>$45,000 to $49,999</td>
<td>$55,000 to $59,999</td>
<td>$50,000 to $54,999</td>
</tr>
<tr>
<td><strong>Total N</strong></td>
<td>1396</td>
<td>1925</td>
<td>3321</td>
</tr>
</tbody>
</table>
Table 2.6: Study 1 – Participants’ Political Party versus their Guess of the Fictional Politician's Political Party

<table>
<thead>
<tr>
<th>Participants' Political Party</th>
<th>Which political party do you think the speaker in the video belongs to?</th>
<th>Dem.</th>
<th>Don't Know / Not Clear</th>
<th>Indep.</th>
<th>Repub.</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Repub.</td>
<td>N</td>
<td>169</td>
<td>149</td>
<td>127</td>
<td>89</td>
<td>534</td>
</tr>
<tr>
<td></td>
<td>%</td>
<td>31.65</td>
<td>27.9</td>
<td>23.78</td>
<td>16.67</td>
<td>100</td>
</tr>
<tr>
<td>Indep.</td>
<td>N</td>
<td>37</td>
<td>68</td>
<td>49</td>
<td>22</td>
<td>176</td>
</tr>
<tr>
<td></td>
<td>%</td>
<td>21.02</td>
<td>38.64</td>
<td>27.84</td>
<td>12.5</td>
<td>100</td>
</tr>
<tr>
<td>Dem.</td>
<td>N</td>
<td>163</td>
<td>189</td>
<td>194</td>
<td>104</td>
<td>650</td>
</tr>
<tr>
<td></td>
<td>%</td>
<td>25.08</td>
<td>29.08</td>
<td>29.85</td>
<td>16</td>
<td>100</td>
</tr>
<tr>
<td>Total</td>
<td>N</td>
<td>369</td>
<td>406</td>
<td>370</td>
<td>215</td>
<td>1,360</td>
</tr>
<tr>
<td></td>
<td>%</td>
<td>27.13</td>
<td>29.85</td>
<td>27.21</td>
<td>15.81</td>
<td>100</td>
</tr>
</tbody>
</table>

Lucid asks for participants' political party before they are able to enroll in any surveys. ANOVA f-statistic=2.84 (p<.0369).
**Figure 2.4**: Study 2 - Impacts of information about deepfakes on disbelief (3-point scale): unconvinced that politician believes what is being said.

The top row illustrates covariate unadjusted mean sentiment in each treatment arm, while the bottom row shows the resulting treatment effects with 95% confidence intervals. The leftmost column illustrates the impact of information about deepfakes added above a deepfake video; the remaining columns illustrate the impact of information about deepfakes added above a real video. All three columns measure impact in terms of the same outcome: how convinced are you that the politician believes what is being said? (3-point scale from 1 = "not at all convinced" to -1 = "very convinced"). The two leftmost treatment effect graphs use Manski bounds to calculate 95% confidence intervals. The rightmost treatment effect graph uses covariate-adjusted standard errors from an OLS regression, as there was no differential attrition.
**Figure 2.5**: Study 2 - Impacts of information about deepfakes on disbelief (binary): don’t know if politician truly believes what is being said

The top row illustrates covariate unadjusted mean sentiment in each treatment arm, while the bottom row shows the resulting treatment effects with 95% confidence intervals. The leftmost column illustrates the impact of information about deepfakes added above a deepfake video; the remaining columns illustrate the impact of information about deepfakes added above a real video. All three columns measure impact in terms of the same outcome: how convinced are you that the politician believes what is being said? (3-point scale from 1="not at all convinced" to -1="very convinced"). The two leftmost treatment effect graphs use Manski bounds to calculate 95% confidence intervals. The rightmost treatment effect graph uses covariate-adjusted standard errors from an OLS regression as there was no differential attrition.
Table 2.7: Study 2 - Hand-coded Romney facts across conditions

<table>
<thead>
<tr>
<th>Category</th>
<th>No Info about Deepfakes</th>
<th>Info about Deepfakes</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Don't Know / No Response Given</td>
<td>28.2%</td>
<td>27.0%</td>
<td>27.6%</td>
</tr>
<tr>
<td>N</td>
<td>818</td>
<td>777</td>
<td>1,595</td>
</tr>
<tr>
<td>Is Republican</td>
<td>16.5%</td>
<td>16.5%</td>
<td>16.5%</td>
</tr>
<tr>
<td>N</td>
<td>477</td>
<td>475</td>
<td>952</td>
</tr>
<tr>
<td>Ran for President</td>
<td>9.7%</td>
<td>10.7%</td>
<td>10.2%</td>
</tr>
<tr>
<td>N</td>
<td>281</td>
<td>309</td>
<td>590</td>
</tr>
<tr>
<td>Subjective Evaluation of Character (e.g., trustworthy, traitor, etc.)</td>
<td>9.2%</td>
<td>9.3%</td>
<td>9.3%</td>
</tr>
<tr>
<td>N</td>
<td>266</td>
<td>268</td>
<td>534</td>
</tr>
<tr>
<td>Is Senator / Former Governor</td>
<td>8.9%</td>
<td>8.3%</td>
<td>8.6%</td>
</tr>
<tr>
<td>N</td>
<td>259</td>
<td>240</td>
<td>499</td>
</tr>
<tr>
<td>Other</td>
<td>8.9%</td>
<td>8.3%</td>
<td>8.6%</td>
</tr>
<tr>
<td>N</td>
<td>258</td>
<td>240</td>
<td>498</td>
</tr>
<tr>
<td>Demographic Fact (e.g., white, male, businessman)</td>
<td>7.7%</td>
<td>8.8%</td>
<td>8.2%</td>
</tr>
<tr>
<td>N</td>
<td>224</td>
<td>252</td>
<td>476</td>
</tr>
<tr>
<td>Mormon</td>
<td>7.6%</td>
<td>7.4%</td>
<td>7.5%</td>
</tr>
<tr>
<td>N</td>
<td>219</td>
<td>214</td>
<td>433</td>
</tr>
<tr>
<td>Is/Was Pro-Choice</td>
<td>1.7%</td>
<td>0.7%</td>
<td>1.2%</td>
</tr>
<tr>
<td>N</td>
<td>49</td>
<td>19</td>
<td>68</td>
</tr>
<tr>
<td>Stance on Other Policies</td>
<td>0.8%</td>
<td>0.9%</td>
<td>0.9%</td>
</tr>
<tr>
<td>N</td>
<td>24</td>
<td>25</td>
<td>49</td>
</tr>
<tr>
<td>Is/Was Pro-Life</td>
<td>0.3%</td>
<td>1.4%</td>
<td>0.8%</td>
</tr>
<tr>
<td>N</td>
<td>9</td>
<td>39</td>
<td>48</td>
</tr>
<tr>
<td>Believe Video is Real</td>
<td>0.3%</td>
<td>0.5%</td>
<td>0.4%</td>
</tr>
<tr>
<td>N</td>
<td>9</td>
<td>15</td>
<td>24</td>
</tr>
<tr>
<td>Believe Video is Fake</td>
<td>0.2%</td>
<td>0.1%</td>
<td>0.2%</td>
</tr>
<tr>
<td>N</td>
<td>5</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>Total</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>N</td>
<td>2,898</td>
<td>2,877</td>
<td>5,775</td>
</tr>
</tbody>
</table>
Table 2.8: Study 1 – Information about deepfakes has no impact on favorability, vote choice, or support for essential oils in medicine (but we are underpowered)

<table>
<thead>
<tr>
<th>Information Treatment Effect</th>
<th>Favor</th>
<th>Vote</th>
<th>More Essential Oils in Medicine?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-0.23</td>
<td>-0.11</td>
<td>-0.26</td>
</tr>
<tr>
<td>95% Upper Bound</td>
<td>0.25</td>
<td>0.08</td>
<td>0.27</td>
</tr>
<tr>
<td>95% Lower Bound</td>
<td>-0.75</td>
<td>-0.32</td>
<td>-0.77</td>
</tr>
</tbody>
</table>

As mentioned in the PAP, this analysis excludes participants who are likely supporters of alternative medicine. This is defined as ALL participants EXCEPT those who BOTH oppose government mandates for vaccination AND support government subsidies for acupuncture procedures.
Table 2.9: Study 1 – Deepfakes have no impact on favorability, vote choice, or support for essential oils in medicine (but we are underpowered)

<table>
<thead>
<tr>
<th>Deepfake Treatment Effect</th>
<th>Favor</th>
<th>Vote</th>
<th>More Essential Oils in Medicine?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.16</td>
<td>0.04</td>
<td>-0.16</td>
</tr>
<tr>
<td>95% Upper Bound</td>
<td>0.64</td>
<td>0.22</td>
<td>0.36</td>
</tr>
<tr>
<td>95% Lower Bound</td>
<td>-0.35</td>
<td>-0.18</td>
<td>-0.66</td>
</tr>
</tbody>
</table>

As mentioned in the PAP, this analysis excludes participants who are likely supporters of alternative medicine. This is defined as ALL participants EXCEPT those who BOTH oppose government mandates for vaccination AND support government subsidies for acupuncture procedures.
Table 2.10: Study 2 - Alternative measure of trust in video as a source of information

<table>
<thead>
<tr>
<th>Information Treatment Effect</th>
<th>Trust TV</th>
<th>Trust Online</th>
</tr>
</thead>
<tbody>
<tr>
<td>95% Upper Bound</td>
<td>0.09</td>
<td>0.05</td>
</tr>
<tr>
<td>95% Lower Bound</td>
<td>-0.07</td>
<td>-0.12</td>
</tr>
</tbody>
</table>

The survey question is “Imagine you saw a video of a politician saying something controversial on [Facebook or Twitter/television news]. How likely are you to believe that the politician actually said what you see in the video?” Answers are on 5-point scale from “Completely certain the video is real” to “Completely certain the video is not real.”

Table 2.11: Study 2 – Republicans appear to have directional (non-significant) increases in favorability when information about deepfakes is added to Romney video

<table>
<thead>
<tr>
<th>Information*Republicanan</th>
<th>Vote</th>
<th>Favor</th>
</tr>
</thead>
<tbody>
<tr>
<td>95% Upper Bound</td>
<td>0.20</td>
<td>0.52</td>
</tr>
<tr>
<td>95% Lower Bound</td>
<td>-0.10</td>
<td>-0.08</td>
</tr>
</tbody>
</table>

Vote is measured via “Given the information that you have, would you consider voting for the politician pictured in the video?” with answers on a 3-point scale: -1 = “No”, 0 = “Don’t Know / Not Sure”, 1 = “Yes.” Favorability is measured via “How favorable or unfavorable is your view of the politician pictured in the video?” with answers on a 7-point scale ranging from extremely unfavorable to extremely favorable.
Table 2.12: Study 2 – Participants who are pro-life appear to have directional (non-significant) increases in favorability when information about deepfakes is added to Romney video

<table>
<thead>
<tr>
<th>Information*Pro-Life</th>
<th>Vote</th>
<th>Favor</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.03</td>
<td>0.12</td>
</tr>
<tr>
<td>95% Upper Bound</td>
<td>0.18</td>
<td>0.43</td>
</tr>
<tr>
<td>95% Lower Bound</td>
<td>-0.12</td>
<td>-0.18</td>
</tr>
</tbody>
</table>

*Vote is measured via “Given the information that you have, would you consider voting for the politician pictured in the video?” with answers on a 3-point scale: -1 = “No”, 0 = “Don’t Know / Not Sure”, 1 = “Yes.” Favorability is measured via “How favorable or unfavorable is your view of the politician pictured in the video?” with answers on a 7-point scale ranging from extremely unfavorable to extremely favorable.*
Table 2.13: Study 2 – No evidence of motivated reasoning

<table>
<thead>
<tr>
<th></th>
<th>Deepfake?</th>
<th>Unconvinced</th>
<th>Don’t Know if Real</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information*Republican</td>
<td>-0.11</td>
<td>-0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>95% Upper Bound</td>
<td>0.17</td>
<td>0.11</td>
<td>0.12</td>
</tr>
<tr>
<td>95% Lower Bound</td>
<td>-0.39</td>
<td>-0.17</td>
<td>-0.06</td>
</tr>
</tbody>
</table>

The three columns represent three outcomes (from left to right): 1) do you believe the video clip is a deepfake? [7-point scale from -3=“strongly disagree” to 3=“strongly agree”]; 2) how convinced are you that the politician believes what is being said? [3-point scale from 1=”not at all convinced” to -1=“very convinced”]; and 3) I don’t know if the politician truly believes what is said [binary variable].
Table 2.14: Study 2 – Some evidence of more distrust immediately before a major election

<table>
<thead>
<tr>
<th>Post Election Indicator</th>
<th>Deepfake?</th>
<th>Unconvinced</th>
<th>Don’t Know if Real</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-0.14</td>
<td>-0.04</td>
<td>-0.01</td>
</tr>
<tr>
<td>95% Upper Bound</td>
<td>0.00</td>
<td>0.03</td>
<td>0.04</td>
</tr>
<tr>
<td>95% Lower Bound</td>
<td>-0.27</td>
<td>-0.11</td>
<td>-0.05</td>
</tr>
</tbody>
</table>

The three columns represent three outcomes (from left to right): 1) do you believe the video clip is a deepfake? [7-point scale from -3=“strongly disagree” to 3=“strongly agree”]; 2) how convinced are you that the politician believes what is being said? [3-point scale from 1=“not at all convinced” to -1=“very convinced”]; and 3) I don’t know if the politician truly believes what is said [binary variable].
Table 2.15: Study 2 – No evidence that the information treatment effect is different before versus after a presidential election

<table>
<thead>
<tr>
<th>Information*Post Election Indicator</th>
<th>Deepfake?</th>
<th>Unconvinced</th>
<th>Don’t Know if Real</th>
</tr>
</thead>
<tbody>
<tr>
<td>95% Upper Bound</td>
<td>0.10</td>
<td>-0.06</td>
<td>0.04</td>
</tr>
<tr>
<td>95% Lower Bound</td>
<td>-0.17</td>
<td>-0.20</td>
<td>-0.05</td>
</tr>
</tbody>
</table>

The three columns represent three outcomes (from left to right): 1) do you believe the video clip is a deepfake? [7-point scale from -3=“strongly disagree” to 3=“strongly agree”]; 2) how convinced are you that the politician believes what is being said? [3-point scale from 1=“not at all convinced” to -1=“very convinced”]; and 3) I don’t know if the politician truly believes what is said [binary variable].
Figure 2.6: Information about deepfakes had no impact on trust in future videos found on TV and online\textsuperscript{21}

The outcome measure has the same wording across both studies: “When you see videos of politicians making controversial statements [on television news/online], how confident are you that they haven’t been faked, altered, or manipulated?” Responses range from 0 = “not at all confident” to 4 = “completely confident.” The two panels illustrate covariate-adjusted treatment effects. Study 1 treatment effects use Manski bounds to calculate 95% confidence intervals. The confidence intervals in Study 2 are estimated using covariate-adjusted standard errors from an OLS regression, as there was no differential attrition. The pooled estimate is estimated using a regression with shared covariates; it excludes all participants in the deepfake condition in Study 2.

\textsuperscript{21} An alternative wording of the same outcome (“Imagine you saw a video of a politician saying something controversial on [television news/Facebook or Twitter]. How likely are you to believe that the politician actually said what you see in the video?”) was used in Study 2 and produced very similar treatment effects. See Table 2.10 in the Supplementary Materials.
2.8.5: Discussion of Secondary Research Question Results

2.8.5.1: Study 1

Due to the necessity of using bounds, we were severely underpowered when examining the secondary research questions in Study 1. We assessed whether information about deepfakes or deepfakes themselves have any impact on favorability towards the politician or the extreme policy (mandating doctors to use essential oils to treat cancer). As shown in Tables 2.8-2.9, we saw no statistically significant differences.

2.8.5.2: Study 2

Study 2 was split into two waves directly before and after the election to assess whether a highly salient election context with high levels of ambient misinformation makes Americans more distrustful of new political information. There is some indication that this may be the case. As seen in Table 2.14 of the Supplementary Materials, we find that participants surveyed before the election have higher levels (0.14 points more) of belief that the video they watched is deepfaked than participants surveyed after the election (p<.05). We emphasize that caution should be used when interpreting these results, as we found that the sample of participants had significantly different demographics (F(34, 1727) = 1.69, p<.01). Even though we control for these differences, it is possible that there are unobserved confounds driving this difference in trust levels.

2.8.5.3: Pooled

Information about deepfakes appears to have some impact on favorability and intention to vote for the politician featured in the video clip. Overall, information about deepfakes decreases intention to vote for the politician by 4.2 percentage points (p<0.05),
with little difference in effect sizes across studies. Similarly, information about deepfakes
decreases the politician’s favorability by 0.17 points on a 7-point scale (from extremely
unfavorable to extremely favorable) (p<0.05). There are, however, strong theoretical reasons
to believe that there should be considerable heterogeneity in these effects. For instance, in
Study 2, pro-life Republicans should be less likely to favor Romney if they believe that he
was or is pro-choice. And since information about deepfakes increases disbelief that Romney
was ever pro-choice, favorability and vote intention levels in the Information condition
should be higher than in the No Information condition. This pattern should be reversed for
pro-choice Democrats. Our results (see Supplementary Materials, Tables 2.11-2.12) are
consistent with this theory in terms of estimate directions but are not statistically significant.

We did not see support for this theory in Study 1—even those participants who
should oppose the extremist essential oils policy reported lower levels of favorability when
treated with information about deepfakes. We must emphasize that we did not have similarly
clear predictors of policy preference as Republicanism for abortion, so we could not
accurately identify people who were for or against extreme policies in favor of holistic
medicine.

2.8.6: Discussion of Motivated Reasoning

The literature on politically-motivated reasoning suggests that people are skeptical of
information that contradicts the beliefs of their political affiliation. If this is the case, we
expect Republicans will be more likely to dismiss the real video of Mitt Romney in Study 2
as a deepfake, because the view expressed by a well-known Republican contradicts the policy
stance of the Republican party. While we do include results illustrating the interaction of the
information treatment and Republican self-identification (see Table 2.13 in the
Supplementary Materials), we cannot directly address the question of motivated reasoning with this data for three reasons. First, Romney had recently become a polarizing figure among Republicans due to his criticism of Donald Trump and support of Black Lives Matter protests. Second, we did not measure Romney favorability pre-treatment. And finally, it is not clear that all rank-and-file Republicans identify being pro-choice as a clearly Republican position; 31% of surveyed Republicans identified themselves as pro-choice in 2019 (Kirzinger et al. 2020).

2.8.7: Ethical Considerations

The authors declare that the human subjects research in this article was reviewed and approved by the Yale University Human Subjects Committee. The authors affirm that this article adheres to the APSA’s Principles and Guidance on Human Subjects Research. Participants were compensated for their participation by the panel provider.

Additionally, these studies were designed such that there was no deception and they avoided ethical gray areas of similar deepfake studies. If we use existing ethical frameworks (e.g., Thomas et al. 2017), the typical risks of an experimental study on deepfakes are 1) the possibility of behavioral/sentiment change (i.e., the possibility of introducing new information into the broader informational environment), 2) the potential for abuse (i.e., results of research can be used by malicious actors) and 3) necessary use (i.e., is it possible to answer the same research question with an alternative design)? We avoid those risks completely, as these studies do not introduce any new misinformation. And because we do not create a deepfake of a real politician, there is no potential for abuse by malicious actors (i.e., there is no material detailing how to create a deepfake of a real politician). Finally, our studies illustrate two alternative designs to study the impact of political deepfakes without
creating a new deepfake of a real politician: 1) using a fictional politician or 2) using a real video of a real politician where the content only *appears* spurious.

### 2.8.8: Materials

#### 2.8.8.1: Study 1 Original PAP

The full PAP can be found online at https://osf.io/rqfz5/?view_only=e2807b367a534262bb6e7aeb5727b999 but is reproduced in full for convenience.

#### 2.8.8.1.1: Hypotheses

**Primary Hypotheses:**

**H1:** Are people are able to discern DeepFakes from real video?

- [DeepFake & NoWarning] compared to [Real & NoWarning] \(\Rightarrow\) Increase in Disbelief

**H2:** How effective are DeepFake warnings?

**H2a:** Do DeepFake warnings cause people to (correctly) believe that DeepFakes are fake?

- [DeepFake & Warning] compared to [DeepFake & NoWarning] \(\Rightarrow\) Increase in Disbelief

**H2b:** Do DeepFake warnings cause people to (incorrectly) believe that real videos are fake?

- [Real & Warning] compared to [Real & NoWarning] \(\Rightarrow\) Increase in Disbelief
H2c: Do DeepFake warnings help people discern DeepFakes from real videos?

- Difference between H2a and H2b → Increase in Disbelief

H3: Do DeepFake warnings make people less likely to believe in any (real or fake) political footage?

- [Warning] compared to [NoWarning] → Decrease in Trust in Footage of Politicians Online
- [Warning] compared to [NoWarning] → No Change in Trust in Footage of Politicians on TV News

**Secondary Hypotheses:**

S_H1a: Are DeepFakes as persuasive as real videos with respect to the favorability/likelihood of voting for the faked politician?

- Among those who are NOT likely to support the expansion of alternative medicine:22

  [DeepFake] compared to [Real] → Increase in Support23

S_H1b: Do warnings reduce the persuasive impact of videos with respect to the favorability/likelihood of voting for the faked politician?

---

22 This population is defined as ALL participants EXCEPT those who BOTH oppose government mandates for vaccination AND support government subsidies for acupuncture procedures.

23 In other words, we expect that people who are against the proposed policy will be more supportive of the politician if they suspect the video is fake.
Among those who are NOT likely to support the expansion of alternative medicine:²⁴ [Warning] compared to [NoWarning] \(\rightarrow\) Increase in Support²⁵

S_H2: Are DeepFakes as persuasive as real videos with respect to the position articulated by the fake politician?

- [DeepFake] compared to [Real] \(\rightarrow\) Decrease in Support for the Expansion of Essential Oils in Medicine

S_H3: Do warnings reduce the persuasive impact of videos with respect to the position articulated by the fake politician?

- [Warning] compared to [NoWarning] \(\rightarrow\) Decrease in Support for the Expansion of Essential Oils in Medicine

2.8.1.1.2: Design Plan

This survey experiment is designed in Qualtrics and is to be implemented via Lucid. All participants who consented, met our eligibility requirements (18+ US citizens who are eligible to vote), and passed the video, sound, and attention checks proceed to the randomization phase. They are randomized to either receive a warning about modified videos or not, and to either view a real video of the speaker or a computer-generated modification thereof. In other words, there are 4 unique treatment conditions:

1) No Warning + Real Video

²⁴ This population is defined as ALL participants EXCEPT those who BOTH oppose government mandates for vaccination AND support government subsidies for acupuncture procedures.

²⁵ In other words, we expect that people who are against the proposed policy will be more supportive of the politician if they suspect the video is fake.
2) Warning + Real Video
3) No Warning + DeepFake Video
4) Warning + DeepFake video

In all 4 conditions, the speaker, a hired actor playing the part of a politician, will speak in favor of a policy position (e.g. mandating the use of essential oils for the treatment of cancer). (Please see “Supplementary Materials: Script” for the exact wording.)

2.8.1.1.3: Sampling Plan

This study will use a sample of individuals paid by Lucid to participate in our Qualtrics survey. We ran a small pilot of this study via Lucid. This pilot was used to devise the analysis code only; the data from this pilot will not be included in the final analysis.

We use the pilot to calculate statistical power. Among our primary hypotheses, we isolated the smallest difference that we believe is meaningfully different from zero. The effect size was 0.1 with a baseline of .45 and a standard deviation of .48. This difference was
between two (of four) treatment arms. To get 80% power, this kind of effect size
necessitates 362 observations in each condition. Since we have 4 conditions in total, we will
need approximately 1,500 observations to detect similarly sized effects.26

2.8.1.1.4: Variables

We use three questions to test H1 and H2:

DeepFake: This video was doctored, manipulated and/or faked by a computer (i.e. it is a
“Deep Fake”).

-3 = Strongly disagree

-2 = Disagree

-1 = Somewhat disagree

0 = Neither agree nor disagree

1 = Somewhat agree

2 = Agree

3 = Strongly agree

NotConvinced:27 How convinced are you that the politician believes what is being said?

1 = Not at all convinced

26 Stata code: sampsi .45 .55, sd(.48) power(.8)

27 To ease exposition, the three variables are coded such that a positive increase in each variable is associated
with more disbelief in the video.
0 = Not sure

-1 = Very convinced

*Don’t Know*: Which statement comes closer to your view?

0 = The politician truly believes what is said in the video

0 = The politician is only saying things in the video in order to get elected

1 = I don’t know if the politician truly believes what is said in the video

To test **H3**, we use:

When you see videos of politicians making controversial statements on television news, how confident are you that they haven’t been faked, altered, or manipulated?

When you see videos of politicians making controversial statements online, how confident are you that they haven’t been faked, altered, or manipulated?

Trust is measured on a 4-point scale (ranging from “0 = not at all confident” to “4 = completely confident”).

To test **S_H1**, we use:

How favorable or unfavorable is your view of the politician pictured in the video?

-3 = Extremely unfavorable

-2 = Unfavorable

-1 = Somewhat unfavorable

0 = Neither favorable nor unfavorable
1 = Somewhat favorable

2 = Favorable

3 = Extremely favorable

Given the information that you have, would you consider voting for the politician pictured in the video?

1 = Yes

-1 = No

0 = Not sure / don’t know

To test S_H2, we use:

The US government should encourage the use of essential oils in medicine

-3 = Strongly disagree

-2 = Disagree

-1 = Somewhat disagree

0 = Neither agree nor disagree

1 = Somewhat agree

2 = Agree

3 = Strongly agree
2.8.1.5: Analysis Plan

For details, please see the attached (annotated) Stata code. The code was written 
before the analysis was conducted.

The analysis will code the variables of interest as depicted above. We exclude only 
those who do not consent, fail to meet our eligibility criteria, or fail the audio, visual, and 
attention checks. We check for differential attrition and check that all pre-treatment 
covariates are balanced across conditions. If there is imbalance on any covariate (p<.05), we 
control for it in our covariate-adjusted specification. For individuals, who did not state where 
they are on the ideological spectrum, we impute their ideology via an OLS regression using 
their self-reported policy stances to predict an ideology score. We include an indicator to 
denote observations that have an imputed ideology.

The main analysis will use a covariate adjusted OLS regression to maximize the 
precision of our estimates. The regression is denoted as SPECIFICATION 3 in the attached 
code and will be run as-is (controlling for ideology, policy stance on vaccination, and policy 
stance on acupuncture) if all pre-treatment covariates are balanced. If there is any imbalance 
(p<.05), we will add the imbalanced covariates to the regression. Unless otherwise specified, 
SPECIFICATION 1 and 2 will be run as robustness checks.

SPECIFICATION 1 will report raw means, while SPECIFICATION 2 is designed 
to better isolate the “scandalous content” component of the treatment. Namely, the policy 
stance is designed to be scandalous regardless of ideology. From the pilot, a non-trivial 
proportion of people who are 1) against government mandates for vaccination and 2) in 
favor of Medicare/Medicaid covering the cost of acupuncture had favorable views of the 
hypothetical politician and were in favor of expanding essential oils in medicine. This may
create two different conceptual treatment constructs: one where the DeepFake is of scandalous content and another where it is of "normal" content—depending on who is viewing the footage. As such, in SPECIFICATION 2, we exclude individuals who are BOTH against mandatory vaccination of school children AND in favor of government subsidies for acupuncture procedures.

Statistical significance for all analyses will be reported at the conventional p<.05 level (2-sided).

To answer S_H1(a-b), we must use only SPECIFICATION 2, since we expect that people who favor the use of essential oils in medicine may be disappointed if they suspect the policy stance has been faked, while people who are against alternative medicine may have the exact opposite effect (they may not judge the politician as harshly if they suspect that the video has been faked). We evaluate only the treatment effects among those who are not highly likely to favor the expansion of essential oils in medicine. We lack sufficient sample size to test for the opposite effect among those who are most likely to favor essential oils in medicine.

As an additional diagnostic, we will also check that the speaker is not perceived as obviously partisan. We will also run a manipulation check with a simple t-test to ensure that the experiment was executed correctly.

2.8.8.2: Study 2 PAP

The full PAP can be found online at https://osf.io/rqfz5/?view_only=e2807b367a534262bb6c7aeb5727b999 but is reproduced in full for convenience.
2.8.8.2.1: Introduction

There is concern that political DeepFakes (AI-doctored videos of politicians speaking) may mislead voters; and there have been several studies evaluating the impact of DeepFake exposure on belief in the content of the faked videos (Wittenberg, Zong, & Rand, 2020; Vaccari & Chadwick, 2020). The study in this pre-analysis plan assesses whether messaging about DeepFakes makes people more likely to disbelieve real video information. Much like warnings about textual fake news caused people to disbelieve real headlines (Clayton et al. 2019), we suspect information about DeepFakes may make people less likely to believe policy statements that are surprising or controversial—even when the voter sees video footage of the politician actually making that statement.

2.8.8.2.2: Design Plan

This survey experiment is designed in Qualtrics and is to be implemented via Lucid. All participants who consented, met our eligibility requirements, and passed the video, sound, and attention checks proceed to the randomization phase. (We discuss exclusions in greater detail in the Exclusion section.) They are randomized to either receive a warning about modified videos or not.

In both conditions, participants view the same historical clip of Romney saying, “And I’ve been very clear on that. I will preserve and protect a woman’s right to choose and I’m devoted and dedicated to honoring my word…” while campaigning for governor of Massachusetts. (Please see the figure below for a screencap of the video.)
2.8.8.2.3: Sampling Plan

This study will use a sample of individuals (expected N of 3,000) paid by Lucid to participate in our Qualtrics survey. We ran a small pilot of this study via Lucid. This pilot was used to devise the analysis code only; the data from this pilot will not be included in the final analysis.

2.8.8.2.4: Research Questions

*Primary Questions:*

(Q1\_believe\_any\_vid) **Do DeepFake warnings make people less likely to believe in any (real or fake) political footage?**

- *Prediction:* Decrease in trust in footage of politicians online
- *Prediction:* Increase in trust in footage of politicians on TV news

(Q2\_disbelieve\_real) **Do DeepFake warnings cause people to (incorrectly) believe that a concrete real video is fake?**

- *Prediction:* Increase in disbelief in veracity of video
• Prediction: Increase in disbelief that Romney was ever pro-choice

**Secondary Questions:**

\[(SQ1\_favorability\_increase)\] Do warnings increase the favorability/likelihood of voting for the faked politician…

\[(SQ1a\_repub\_fav\_increase)\] more among Republicans than among Democrats?

• Prediction: The interaction of treatment and a Republican indicator is expected to be positive

\[(SQ1b\_pro\_life\_fav\_increase)\] more among pro-lifers than among pro-choicers?

• Prediction: The interaction of treatment and a pro-life indicator is expected to be positive

\[(SQ2\_motiv\_reason)\] Are Republicans more likely to disbelieve the real video than Democrats (e.g., motivated reasoning)?

• Prediction: For the outcomes 1) disbelief in veracity of video and 2) disbelief that Romney was ever pro-choice…

  o the interaction of treatment and a Republican indicator is expected to be positive

\[(SQ3\_election1)\] Is distrust stronger directly before the 2020 US General election?

• Prediction: For the outcome disbelief in veracity of video…

  o an indicator for whether the respondent took the survey after midnight on November 3rd, 2020 is expected to be negative
Are effects on disbelief stronger directly before the 2020 US General election?

- Prediction: For the outcome of disbelief in veracity of video…
  - the interaction of treatment and an indicator for whether the respondent took the survey after midnight on November 3rd, 2020 is expected to be negative

2.8.8.2.5: Outcome Measures

We use the following survey questions to evaluate \( Q1_{\text{believe-any-vid}} \):

*Trust Faked:* When you see videos of politicians making controversial statements [on television news/online], how confident are you that they haven't been faked, altered, or manipulated?

- 0 = not at all confident
- 1 = somewhat confident
- 2 = very confident
- 3 = completely confident

*Trust Actually Said:* Imagine you saw a video of a politician saying something controversial on [television news/Facebook or Twitter]. How likely are you to believe that the politician actually said what you see in the video?

- 0 = Completely certain the video is not real
- 1 = Somewhat certain the video is not real
- 2 = Don't know if the video is real or not
3 = Somewhat certain the video is real

4 = Completely certain the video is real

For $Q2\_disbelieve\_real$ and $SQ2\_motiv\_reason$ we use the following questions to measure disbelief\textsuperscript{28} in veracity:

**DeepFake:** This video was doctored, manipulated and/or faked by a computer (i.e. it is a “Deep Fake”).

-3 = Strongly disagree

-2 = Disagree

-1 = Somewhat disagree

0 = Neither agree nor disagree

1 = Somewhat agree

2 = Agree

3 = Strongly agree

**NotConvinced:** How convinced are you that the politician believes what is being said?

1 = Not at all convinced

0 = Not sure

\textsuperscript{28} To ease exposition, the three variables are coded such that a positive increase in each variable is associated with more disbelief in the video.
-1 = Very convinced

Don’t Know: Which statement comes closer to your view?

0 = The politician truly believes what is said in the video

0 = The politician is only saying things in the video in order to get elected

1 = I don’t know if the politician truly believes what is said in the video

and the following two questions (Free Response and Fact List) will measure disbelief in the content:

Free Response: In the text boxes below, please write three facts you know to be true about Mitt Romney. If you can’t think of a fact, you may answer “don’t know” in as many boxes as you need to.

[Any answer containing “choose”, “choice”, “abort”, and “life” will be automatically flagged as =0 (=1 otherwise) and then will be manually reviewed to confirm that the free-response correctly identifies that Romney was once pro-choice.]

Fact List In the following table, please mark all positions that you believe Mitt Romney has ever held.

[Facts about Romney]

Supported women’s access to abortion

0 = True

1 = False
1 = Not sure / Don’t know

To test $SQ1_{favorability\_increase}$, we use:

How favorable or unfavorable is your view of the politician pictured in the video?

-3 = Extremely unfavorable
-2 = Unfavorable
-1 = Somewhat unfavorable
0 = Neither favorable nor unfavorable
1 = Somewhat favorable
2 = Favorable
3 = Extremely favorable

Given the information that you have, would you consider voting for the politician pictured in the video?

1 = Yes
-1 = No
0 = Not sure / don’t know

2.8.8.2.6: Analysis

The main analysis will use a covariate adjusted OLS regression. The covariates are: Lucid demographic variables, policy stances, ideology, Big Five personality traits, political knowledge, and political interest. For details, please see the attached annotated Stata code. The code was written before the analysis was conducted. The analysis will code the variables
of interest as depicted above. We will also check for covariate balance using a logistic regression (see line 22 of the accompanying Stata code).

2.8.8.2.7: Exploratory Analysis

As an additional exploratory analysis, we conduct the same analysis as in SQ2_motiv_reason on the following outcome variables: believing that Romney was once the governor of Massachusetts and is currently pro-life. We expect these treatment effects to have the same direction as in SQ2_motiv_reason but smaller in magnitude.

Finally, we will check to see if a participant’s personal stance on abortion changes as a result of the Warning. We do not expect that there will be much of an effect, but any individual with weak views on this issue may opt to adopt elite cues. We expect that the pre-post difference of participants’ own stance on abortion will be larger in favor of pro-choice in the No Warning condition than in the Warning condition.

2.8.8.2.8: Missing Data

For individuals, who did not state where they are on the ideological spectrum, we will impute their ideology via an OLS regression using their self-reported policy stances to predict an ideology score. We will include an indicator to denote observations that have an imputed ideology.

Since there may be individuals who are randomized but do not finish the survey (and thus do not provide outcome responses), we will check for differential attrition using a chi-squared test (see line 13 in the Stata code). If there is no differential attrition, we conclude that missingness occurs at random and exclude all affected observations from analysis. If there is differential attrition (i.e., chi-squared p-value<.05), we will use inverse probability weights to weight the population of all subjects who passed the attention checks. The
missingness propensity score will be fit via a logistic regression of missingness on covariates, treatment, and their interactions. Any propensity scores under .05 will be Winsorized to 0.05.

2.8.8.2.9: Exclusions

We exclude all participants who satisfy any one of the following conditions:

- Did not consent or meet our eligibility criteria (18+ US citizens who are eligible to vote).
- Have duplicate IP addresses and reached randomization.
- Failed any one of the audio, visual, and attention checks.
- Self-reported age was not within one year of their age according to Lucid.

All main analyses will also exclude those participants who were not able to successfully identify Mitt Romney’s political party.

For $SQ3_{election1}$ and $SQ4_{election2}$, we exclude respondents who:

- Take any portion of the survey between 8am and midnight November 3rd

2.8.8.2.10: Robustness and Placebo Checks

Romney’s pro-choice stance in the video is atypical of Republicans but not Democrats. If the participant does not know Romney’s political party, the content of the

---


30 This is due to data quality concerns raised by Aronow, P. M., Kalla, J., Orr, L., & Ternovski, J. (2020). Evidence of Rising Rates of Inattentiveness on Lucid in 2020. https://doi.org/10.31235/osf.io/8sbe4

31 We interact knowledge of Romney’s partisanship with our treatment effects as a robustness check below.
video may not seem surprising. Responses from these participants will be tested as a robustness check with the expectation that all Romney-specific treatment effects (i.e., all questions with the exception of $Q1\_believe\_any\_vid$) will be reduced among individuals who were not able to correctly identify Romney’s party. This will be evaluated by interacting an indicator variable (where correctly identifying Romney’s party $= 1$, $= 0$ otherwise) with treatment.

We will use the same specification as $SQ2\_motiv\_reason$ on the Romney facts not mentioned above as a placebo check (i.e., we expect to see no treatment effects on Romney facts that are unrelated to the content of the video).

2.8.8.3: Study 1 Survey Script

The full script is titled DF_Study1_Script.pdf and can be found online at https://osf.io/rqfz5/?view_only=e2807b367a534262bb6c7aeb5727b999 but is provided here in full for convenience:

---
Start of Block: Eligibility Criteria

Q46 Are you 18 years of age or older?

- Yes (1)
- No (2)

---
Q47 Are you a citizen of the US?

- Yes (1)
- No (2)

Q48 Are eligible to vote in the US?

- Yes (1)
- No (2)

End of Block: Intro Continued

Start of Block: AUDIOVISUAL CHECK

Q57 In this survey, you will watch a brief political speech. Before showing you the speech, we want to make sure videos properly play for you. Please watch this short clip. Turn up your volume and when you are ready, click the play button to start.

[MOBILE ONLY] If the video appears cut off, please rotate your phone.

Page Break

Q58 How many fingers did the actor hold up?

Q59 What is the number the speaker said?
Q60 For our research, careful attention to survey questions is critical! To show that you are paying attention please select "I have a question."

○ I understand (1)

○ I do not understand (2)

○ I have a question (3)

End of Block: AUDIOVISUAL CHECK

Start of Block: Ideology (ANES)

Q9 When it comes to politics do you usually think of yourself as extremely liberal, liberal, slightly liberal, moderate or middle of the road, slightly conservative, conservative, extremely conservative, or haven't you thought much about this?

○ Extremely liberal (1)

○ Liberal (2)

○ Slightly liberal (3)

○ Moderate or middle of the road (4)

○ Slightly conservative (5)

○ Conservative (6)

○ Extremely conservative (7)

○ Other (8) ________________________________________________

○ I haven't thought much about this (9)

-----------------------------------------------------------------
Q10 People are very busy these days and many do not have time to follow what goes on in the government. We are testing whether people read questions. To show that you've read this much, answer both "extremely interested" and "slightly interested".

☐ Extremely interested (1)

☐ Very interested (2)

☐ Moderately interested (3)

☐ Slightly interested (4)

☐ Not interested at all (5)

End of Block: Ideology (ANES)

Start of Block: Policy Preferences (from Broockman, 2016)
Q11 Do you agree or disagree with the following statements? (Please pick the option that most accurately represents your views.)

<table>
<thead>
<tr>
<th>Statement</th>
<th>Agree (1)</th>
<th>Disagree (2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>The federal government should pay for medical care for elderly Americans.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Legalize the recreational use of marijuana.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Increase taxes for those making over $250,000 per year.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Women should have a constitutional right to have abortions.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Same-sex couples should be allowed to marry.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Implement a universal healthcare program to guarantee coverage to all</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Americans, regardless of income.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>There should be strong restrictions on the purchase and possession of</td>
<td></td>
<td></td>
</tr>
<tr>
<td>guns.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Illegal immigrants should not be allowed to enroll in government food</td>
<td></td>
<td></td>
</tr>
<tr>
<td>stamp programs.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Q28 Do you agree or disagree with the following statements? (Please pick the option that most accurately represents your views.)

<table>
<thead>
<tr>
<th>Statement</th>
<th>Agree (1)</th>
<th>Disagree (2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Include sexual orientation in federal anti-discrimination laws. (4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prohibit the use of affirmative action by state colleges and universities. (5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I support the death penalty in my state. (6)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prohibit the EPA from regulating greenhouse gas emissions. (7)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Health insurance plans should be required to fully cover the cost of birth control. (8)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>The federal government should subsidize student loans for low income students. (9)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medicare and Medicaid should include coverage for acupuncture. (10)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All students should be required to be vaccinated before starting school. (11)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

End of Block: Policy Preferences (from Broockman, 2016)
Q17
On the next page, you will watch a brief political speech. You will then be asked about how you feel about the person making the speech and whether or not you would vote for them in the upcoming election. Please keep your volume turned up.

[Randomization and Video]

Start of Block: Outcomes

Q19 How favorable or unfavorable is your view of the politician pictured in the video?

- Extremely unfavorable (1)
- Unfavorable (4)
- Somewhat unfavorable (5)
- Neither favorable nor unfavorable (6)
- Somewhat favorable (7)
- Favorable (8)
- Extremely favorable (9)
Q20 Given the information that you have, would you consider voting for the politician pictured in the video?

- Yes (1)
- No (2)
- Not sure / don't know (4)
Q21 Rank how much you believe the politician agrees with each of the following statements.

<table>
<thead>
<tr>
<th></th>
<th>Strongly disagrees with (1)</th>
<th>Somewhat disagrees with (2)</th>
<th>Not Sure (3)</th>
<th>Somewhat agrees with (4)</th>
<th>Strongly agrees with (5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gas prices are too high (1)</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>The US government should encourage the use of essential oils in medicine. (2)</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
</tbody>
</table>

Q22 How convinced are you that the politician believes what is being said?

- ○ Not at all convinced (1)
- ○ Not sure (2)
- ○ Very convinced (3)
Q23 Which statement comes closer to your view?

○ The politician truly believes what is being said in the video. (1)

○ The politician is only saying things in the video in order to get elected. (2)

○ I don't know if the politician truly believes what is said in the video. (3)
Q24 How trustworthy will you find videos you see on television news of politicians making controversial statements?

○ Not at all trustworthy (1)

○ Somewhat trustworthy (2)

○ Very trustworthy (3)

○ Completely trustworthy (4)

Q25 How trustworthy will you find videos you see online of politicians making controversial statements?

○ Not at all trustworthy (1)

○ Somewhat trustworthy (2)

○ Very trustworthy (3)

○ Completely trustworthy (4)
Q26 Please rank how much **you** agree or disagree with the following statements.

<table>
<thead>
<tr>
<th>Strongly disagree (1)</th>
<th>Disagree (2)</th>
<th>Somewhat disagree (3)</th>
<th>Neither agree nor disagree (4)</th>
<th>Somewhat agree (5)</th>
<th>Agree (6)</th>
<th>Strongly agree (7)</th>
</tr>
</thead>
<tbody>
<tr>
<td>This video was doctored, manipulated and/or faked by a computer (i.e. it is a &quot;Deep Fake&quot;). (1)</td>
<td>☐ ☐ ☐ ☐ ☐ ☐ ☐</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The US government should encourage the use of essential oils in medicine. (2)</td>
<td>☐ ☐ ☐ ☐ ☐ ☐ ☐</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Q27 Which political party do you think the speaker in the video belongs to?

- Democrat (1)
- Republican (2)
- Independent / Other (3)
- Don't Know / Unclear (4)

End of Block: Outcomes

2.8.8.4: Study 2 Survey Script

The full script is titled Study 2 - Script.pdf and can be found online at https://osf.io/rqfz5/?view_only=e2807b367a534262bb6c7aeb5727b999 but is provided here in full for convenience:

Start of Block: Eligibility Criteria

Q46 Are you 18 years of age or older?

- Yes (1)
- No (2)

Q47 Are you a citizen of the US?

- Yes (1)
- No (2)
Q48 Are you eligible to vote in the US?

- Yes (1)
- No (2)

Q65 How familiar are you with the politician Mitt Romney?

- Very familiar (1)
- Somewhat familiar (2)
- Never heard of (3)

Q67 Which political party is Mitt Romney affiliated with? Some people don't know politicians' political affiliations. If you're not sure, please select not sure / don't know.

- Democrat (1)
- Republican (2)
- Independent / Other (3)
- Not Sure / Don't Know (4)
Q10 People are very busy these days and many do not have time to follow what goes on in the government. We are testing whether people read questions. To show that you've read this much, answer both "extremely interested" and "slightly interested".

☐ Extremely interested (1)

☐ Very interested (2)

☐ Moderately interested (3)

☐ Slightly interested (4)

☐ Not interested at all (5)

End of Block: Intro Continued

Start of Block: AUDIOVISUAL CHECK

Q57 In this survey, you will watch a brief political speech. Before showing you the speech, we want to make sure videos properly play for you. Please watch this short clip. Turn up your volume and when you are ready, click the play button to start.

[IF MOBILE] If the video appears cut off, please rotate your phone.

Page Break

Q58 How many fingers did the actor hold up?
Q59 What is the number the speaker said?

________________________________________

Q60 For our research, careful attention to survey questions is critical! To show that you are paying attention please select "I have a question."

  o I understand (1)
  o I do not understand (2)
  o I have a question (3)

End of Block: AUDIOVISUAL CHECK

Start of Block: Ideology (ANES)

Q9 When it comes to politics do you usually think of yourself as extremely liberal, liberal, slightly liberal, moderate or middle of the road, slightly conservative, conservative, extremely conservative, or haven't you thought much about this?

  o Extremely liberal (1)
  o Liberal (2)
  o Slightly liberal (3)
  o Moderate or middle of the road (4)
  o Slightly conservative (5)
  o Conservative (6)
  o Extremely conservative (7)
  o Other (8) ________________________________
  o I haven't thought much about this (9)
Q11 Do you agree or disagree with the following statements? (Please pick the option that most accurately represents your views.)

<table>
<thead>
<tr>
<th>Statement</th>
<th>Agree (1)</th>
<th>Disagree (2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>The federal government should pay for medical care for elderly Americans. (36)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Legalize the recreational use of marijuana. (37)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Increase taxes for those making over $250,000 per year. (38)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Women should have a constitutional right to have abortions. (39)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Same-sex couples should be allowed to marry. (40)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Implement a universal healthcare program to guarantee coverage to all Americans, regardless of income. (41)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>There should be strong restrictions on the purchase and possession of guns. (42)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Illegal immigrants should not be allowed to enroll in government food stamp programs. (43)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Q28 | Do you agree or disagree with the following statements? (Please pick the option that most accurately represents your views.)

<table>
<thead>
<tr>
<th>Statement</th>
<th>Agree (1)</th>
<th>Disagree (2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Include sexual orientation in federal anti-discrimination laws. (4)</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Prohibit the use of affirmative action by state colleges and universities. (5)</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>I support the death penalty in my state. (6)</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Prohibit the EPA from regulating greenhouse gas emissions. (7)</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Health insurance plans should be required to fully cover the cost of birth control. (8)</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>The federal government should subsidize student loans for low income students. (9)</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Medicare and Medicaid should include coverage for acupuncture. (10)</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>All students should be required to be vaccinated before starting school. (11)</td>
<td>☐</td>
<td>☐</td>
</tr>
</tbody>
</table>

End of Block: Policy Preferences (from Broockman, 2016)
Q106 We will now ask you some questions about yourself.

I see myself as someone who...

<table>
<thead>
<tr>
<th></th>
<th>Strongly Agree (1)</th>
<th>Agree (2)</th>
<th>Neither Agree nor Disagree (3)</th>
<th>Disagree (4)</th>
<th>Strongly Disagree (5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>is reserved (1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>is generally trusting (2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>does a thorough job (3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>is relaxed, handles stress well (4)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>has an active imagination (5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>is outgoing, sociable (6)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tends to find fault in others (7)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tends to be lazy (8)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gets nervous easily (9)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>has few artistic interests (10)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

End of Block: Big5

On the next page, you will watch a brief political speech. You will then be asked about how
you feel about the person making the speech and whether or not you would vote for them in the upcoming election. Please keep your volume turned up.

[IF MOBILE:] If the video appears cut off, please rotate your phone.

[RANDOMIZATION AND VIDEO]

Start of Block: Outcomes1

Q19 How favorable or unfavorable is your view of the politician pictured in the video?

○ Extremely favorable (1)

○ Favorable (4)

○ Somewhat favorable (5)

○ Neither favorable nor unfavorable (6)

○ Somewhat unfavorable (7)

○ Unfavorable (8)

○ Extremely unfavorable (9)
Q20 Given the information that you have, would you consider voting for the politician pictured in the video?

- Yes (1)
- No (2)
- Not sure / don't know (4)
Q22 How convinced are you that the politician believes what is being said?

○ Very convinced (3)
○ Not sure (2)
○ Not at all convinced (1)

Q23 Which statement comes closer to your view?

○ The politician truly believes what is being said in the video. (1)
○ I don't know if the politician truly believes what is said in the video. (3)
○ The politician is only saying things in the video in order to get elected. (2)
Q89 When you see videos of politicians making controversial statements on television news, how confident are you that they haven't been faked, altered, or manipulated?

- Completely confident (4)
- Very confident (3)
- Somewhat confident (2)
- Not at all confident (1)
Q90 When you see videos of politicians making controversial statements **online**, how confident are you that they haven't been faked, altered, or manipulated?

- [ ] Completely confident (4)
- [ ] Very confident (3)
- [ ] Somewhat confident (2)
- [ ] Not at all confident (1)
Q91 Imagine you saw a video of a politician saying something controversial on television news. How likely are you to believe that the politician actually said what you see in the video?

- Completely certain the video is real (1)
- Somewhat certain the video is real (2)
- Don't know if the video is real or not (3)
- Somewhat certain the video is not real (4)
- Completely certain the video is not real (5)
Q92 Imagine you saw a video of a politician saying something controversial on Facebook or Twitter. How likely are you to believe that the politician actually said what you see in the video?

- Completely certain the video is real (1)
- Somewhat certain the video is real (2)
- Don't know if the video is real or not (3)
- Somewhat certain the video is *not* real (4)
- Completely certain the video is *not* real (5)

End of Block: Outcomes2
Q104 In the following table, please mark all facts you know are **currently true** about the politician, **Mitt Romney**.

<table>
<thead>
<tr>
<th>Fact</th>
<th>True (1)</th>
<th>False (2)</th>
<th>Not sure (3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is Mormon (1)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Is Republican (2)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Currently Senator of Utah (3)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wants to cut taxes (4)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Opposes marijuana legalization (5)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Opposes women's access to abortion (6)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Opposes same-sex marriage (7)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Q105 In the following table, please mark all facts you know have been true in the past about the politician, Mitt Romney.

<table>
<thead>
<tr>
<th></th>
<th>True (1)</th>
<th>False (2)</th>
<th>Not sure (3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Was Baptist (1)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Was a Democrat (2)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Was Governor of Massachusetts (3)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Raised taxes (4)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Supported marijuana legalization (5)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Supported women's access to abortion (6)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Supported same-sex marriage (7)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Q106 Please rank how much you agree or disagree with the following statements.

<table>
<thead>
<tr>
<th>Strongly agree (1)</th>
<th>Agree (2)</th>
<th>Somewhat agree (3)</th>
<th>Neither agree nor disagree (4)</th>
<th>Somewhat disagree (5)</th>
<th>Disagree (6)</th>
<th>Strongly disagree (7)</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="circle1" alt="Circle" /></td>
<td><img src="circle2" alt="Circle" /></td>
<td><img src="circle3" alt="Circle" /></td>
<td><img src="circle4" alt="Circle" /></td>
<td><img src="circle5" alt="Circle" /></td>
<td><img src="circle6" alt="Circle" /></td>
<td><img src="circle7" alt="Circle" /></td>
</tr>
<tr>
<td>The video you watched in this survey was doctored, manipulated and/or faked by a computer (i.e. it is a &quot;Deep Fake&quot;). (1)</td>
<td><img src="circle1" alt="Circle" /></td>
<td><img src="circle2" alt="Circle" /></td>
<td><img src="circle3" alt="Circle" /></td>
<td><img src="circle4" alt="Circle" /></td>
<td><img src="circle5" alt="Circle" /></td>
<td><img src="circle6" alt="Circle" /></td>
</tr>
<tr>
<td>Women should have a constitutional right to have abortions. (4)</td>
<td><img src="circle1" alt="Circle" /></td>
<td><img src="circle2" alt="Circle" /></td>
<td><img src="circle3" alt="Circle" /></td>
<td><img src="circle4" alt="Circle" /></td>
<td><img src="circle5" alt="Circle" /></td>
<td><img src="circle6" alt="Circle" /></td>
</tr>
</tbody>
</table>

End of Block: Outcomes3
2.8.8.5: Study 1 Video Transcript

ACTOR: I believe that the future of health care is going to involve a mix of old and new methods of treatment. Despite what the pharmaceutical companies tell you, it turns out that not all medical professionals think that the best way to treat cancer is to simply zap bodies with radiation and hand out more pills. Natural and organic essential oils are safer, more affordable, and more effective at treating and even curing cancer than the drugs sold by pharmaceutical companies. If I am elected, the first thing I will do is work to pass a law requiring doctors to use essential oils to treat cancer. Only if essential oils don’t work will they be allowed to use pharmaceutical drugs, radiation, and conventional medicine. I believe that this will save the American healthcare system millions of dollars every year and save countless lives.

2.8.8.6: Final Study 1 Data Cleanup Code

The final code used for data cleanup is titled study1_data_cleanup_FINAL.do and can be found online at https://osf.io/rqfz5/?view_only=e2807b367a534262bb6c7aeb5727b999

Please note that it has slight modifications from the original code (STUDY1_1exclusions_recode_cleanup_v2.do), which are described fully in Section 2 of the Supplementary Materials, Deviations from Pre-Analysis Plan.

2.8.8.7: Final Study 2 Data Cleanup Code

The final code used for data cleanup is titled study2_data_cleanup_FINAL.do and can be found online at https://osf.io/rqfz5/?view_only=e2807b367a534262bb6c7aeb5727b999

Please note that it has slight modifications from the original code (STUDY_2exclusions_recode_clean.do), which are described fully in Section 2 of the Supplementary Materials, Deviations from Pre-Analysis Plan.
3. The Impact of Trump Tweets and 2016 Victories on Hate Music Listenership

3.1: Abstract

Far-right music serves an important function in organized hate groups worldwide; one of its primary uses is recruitment. In this paper, I identified frequent listeners of hate music on the music listenership website, Last.fm, and tracked their song plays and the song plays of their Last.fm friends before and after a set of unanticipated Trump-related events between 2015 and 2017 (which include presidential primary victories and Trump posting xenophobic content on Twitter). I found that friends of frequent listeners of hate music, who themselves have not previously listened to hate music, experienced increases in hate music listenership after Trump-related events. These effects appear to be larger for friends of hate music listeners who are more influential (as measured by eigenvector centrality) in the hate music listener social network. These results suggest that Trump-related events are impetuses for recruiting and social influence by users who are likely committed to far-right ideologies and their efforts appear to be successful.

3.2: Introduction

In the fall of 2020, members of a right-wing militia were arrested for, in the words of the FBI, plotting “the violent overthrow of certain government and law-enforcement
components” (Sheth and Haltiwanger, 2020; p. 1). Only months later, in January 2021, the US capitol building was stormed, with many of the rioters affiliated with far-right groups including “self-described Nazis and white supremacists” (Diaz & Treisman, 2021; p. 1). Former president Donald Trump has been impeached for this “incitement of insurrection” (Naylor, 2021).

Empirical research has linked former president Donald Trump’s 2016 general election victory with increases in misogyny (e.g., Huang and Low, 2017), xenophobia (e.g., Crandall, Miller & White, 2018; Bursztyn, Egorov, Fiorin, 2019), and racism (Giani & Meon, 2019); exposure to statements made by Trump in survey experiments caused increases in racist sentiments and xenophobia (e.g., Schaffner, 2020; Newman et al. 2021). These increases in bigotry have been dubbed by some as the “Trump effect” (Costello 2016). But this research has largely focused on the interaction between elected politicians and typical American voters, which, while important, does not address the influence of a third strategic actor: organized hate groups (such as the American Nazi Party). Including hate groups in this interaction is particularly important, because they often not only endorse political violence but have been empirically linked to ideologically-motivated fatal violence in counties where they operate (Adamczyk et al., 2014). Outside of the US, communications from far-right extremists have been modelled as part of the causal pathway that lead to hate crimes (Dancygier et al. 2021). The FBI recently raised the threat level of domestic far-right hate groups to the same level as the international terrorist group, ISIS (Woodward, 2020). As such, it is critical to understand how hate group recruitment efforts interact with xenophobic and racist rhetoric from a prominent elected politician. Particularly, does a political elite’s normalization of racism and xenophobia go further than activate casual racism—does it play a part in the radicalization of Americans into far-right ideologies?
Anthropological and sociologists have found that far-right groups make extensive use of hate music (e.g., bands with such names as “Aryan Terrorism”) for recruitment, organizing, and expression of racist beliefs (e.g., Messner et al., 2007; Corte, & Edwards, 2008; Shekhovtsov, 2013; Woolf & Hulsizer, 2004). Indeed, as I discuss in subsequent sections, far-right music can be construed as a proxy for subscription to extreme far-right ideologies that go far beyond casual racism. As such, I generated a dataset of over half a billion song plays from ~250,000 users on the website Last.fm, one of the largest public repositories for music listenership in the world. I then applied an interrupted time series estimator to examine listenership behavior of Last.fm users before and after various, unanticipated Trump-related events. I find that for Last.fm users who have not listened to hate music before but are friends with top listeners of hate music, hate music listenership increased after unanticipated Trump-related events. This suggests that there is a link between Trump and the activation of far-right ideologies among music listeners in the US.

This paper is organized as follows. First, I discuss the motivation behind this research question and review prior studies on the relationship between political elites, hate groups, and far-right radicalization. I then describe my theoretical framework. Next, I describe my data and the empirical methods used. Finally, I present the results and conclude.

### 3.3: Motivation and Prior Research

Political elites’ influence on voter attitudes and behavior is a well-studied topic in political science. While the initial formulations of the theoretical framework for democracy understandably focused on how voters with static policy preferences affect elected politicians’ behaviors (e.g., Black, 1948), there has since been extensive research into how political elites themselves affect the policy preferences and political attitudes of voters (e.g.,
Abramowitz 1978; Gabel and Scheve 2007; Lenz 2009, 2012; Minozzi et al. 2015; Broockman and Butler, 2017; Barber & Pope, 2019; Agadjanian, 2021). Just prior to Donald Trump’s presidency, this research focused on how political elites exploited subtle cues to prime racial resentment among whites; one of the key takeaways is that only subtle, “implicit” cues seemed to consistently work (e.g., Stephens-Dougan 2016; White, 2007). But the anti-prejudice norms that underpinned these effects appeared to change dramatically in recent years (Valentino, Neuner, & Vandenbroek 2018). After Donald Trump’s victories, scholars looked at the impact of explicit racial and xenophobic statements made by Trump on typical voters’ attitudes towards people of color and immigrants (e.g., Schaffner, 2020; Newman et al. 2021). Schaffner (2020) found that exposure to an elected elite flouting anti-prejudice norms made the participants of a survey experiment less likely to endorse norms against the expression of prejudice.

This research is important in showing direct effects of elites’ prejudiced rhetoric on highly consequential voter attitudes; for instance, such attitudes may ultimately swing elections in favor of politicians who implement prejudiced policies (for a more extensive discussion see Schaffner 2020). But these studies do not address an important intermediary: organized hate groups. After all, it isn’t just politicians and political elites that influence public opinion. In certain contexts, organized interest groups can have large effects in reducing prejudice (e.g., Broockman and Kalla, 2016; Kalla and Broockman, 2020), but political scientists have also noted the many ways they can increase it. For instance, Green, Glaser & Rich (1998) provide historical examples illustrating that it is both “political elites and organizations [that can attribute] blame and [foment] public resentment toward minority groups in times of economic contraction” (p. 89). They draw on the findings of Foner (1975), who found unions and elected politicians alike argued that emancipated black citizens
would threaten white workers and that such propaganda may have led to the 1917 St. Louis riots (Green, Glaser & Rich, 1998). An in-depth analysis of the Ku Klux Klan, meanwhile, noted the group’s strategy of exploiting economic conditions for recruitment and persuasion efforts (Wade 1987; Green, Glaser & Rich 1998). If organized groups such as the KKK exploit economic conditions, it is likely that they may want to exploit political conditions, as well.32

3.3.1: Hate Groups

Before discussing the interaction of hate groups, political elites, and the electorate, it is necessary to define far-right, hate groups more precisely. I use the terms far-right groups and hate groups interchangeably, as my definition of hate group follows Fording & Cotter (2014), who define the construct “as any group associated with racial or ethnic hatred…[and] is associated with the ‘white racist right wing’” (p. 2-3). While there are hate groups that are outside of this definition, for instance, SLPC classifies black separatist groups as hate groups, the majority of far-right groups in the US share some form of commitment to white supremacy (SLPC, 2020). White supremacist groups may have differences but all espouse the myth of “white genocide” and “envision a racially exclusive world where ‘nonwhites’ are vanquished, segregated, or at least subordinated to Ayran authority.” (Futrell, Simi, &

---

32 This gap in the literature may be partially motivated by the waning influence of infamous hate groups such as the Ku Klux Klan. A recent Time headline read “9 People Showed Up for a KKK Rally in Dayton, Ohio. They Were Drowned Out by 600 Protestors” (Law 2019) and recent Southern Law Poverty Center (SLPC) reports confirmed the decline of the Ku Klux Klan (SLPC 2019). However, as SLPC notes, the decline of older hate groups does not necessarily translate to an overall decline in white supremacist hate group membership in America (SLPC 2019).
Gottschalk, 2006; p. 281). Unsurprisingly, hate groups have been empirically linked to ideologically-motivated fatal violence in counties where they operate (Adamczyk et al., 2014).

White-supremacist hate groups share another trait—far-right music scenes are often essential to their functioning and growth (e.g., Messner et al., 2007; Corte, & Edwards, 2008; Shekhovtsov, 2013; Woolf & Hulsizer, 2004). Far-right music is often used to recruit new members, especially youths (e.g., Corte & Edwards, 2008). For instance, “[i]n 2004, Panzerfaust Records, a successful North American White Power label, drew on a network of volunteers to help distribute a sample CD to middle and high school students across the country” (Corte & Edwards, 2008, p. 14). Far-right music artists are also often directly involved with political and para-military organizations (Shekhvotsov, 2013). One of the clearest examples in recent news is that the organizers of the largest far-right music festival in Europe were also members of the far-right Azov movement (Hume, 2019), which gained notoriety in 2014 due to the alleged war crimes of affiliated militia groups like the neo-Nazi Azov Battalion (Walker, 2014; Sharkov, 2014).

Far-right music spans many genres but tends to share one characteristic: the lyrical themes, which include “Aryan nationalism, white power, race war, anti-Semitism, anti-immigration, anti-race-mixing, and white victimization” (Futrell, Simi, & Gottschalk, 2006; p. 281). Even the band names are often explicit in their ideological message (e.g., Aryan Terrorism, Racial Purity). Indeed, scholars of far-right violence find that “[g]roups that recruit at… concerts… are likely to be attracting members that are more prone to participate in violence” (Chermak, Freilich, & Suttmoeller, 2013 p. 196). As such, consumption of far-right music may be construed as a proxy for subscription to extreme far-right ideologies that go far beyond casual racism.
3.3.2: The Relationship between Hate Groups and Trump

Non-profits like the Southern Law Poverty Center (SLPC) have connected Trump’s xenophobic rhetoric to the growth of far-right hate groups (e.g., Beirich & Buchanan, 2018), however their analysis focused primarily on numbers of hate groups and not concrete estimates of total membership. Far-right hate groups are notoriously ephemeral with the majority of the organizations dissolving in less than a year of existence (Chermak, Freilich, & Suttmoeller, 2013). Other scholars did find that increases in hate speech online after Trump’s general election victory (e.g., Scrivens et al., 2020; Zannettou et al., 2020). But increases in socially adverse sentiments is distinct from subscription to the ideologies of organized far-right groups. It is also possible that individuals who are already dedicated members of far-right groups are simply more vocal and more mobilized (e.g., make duplicate user accounts), but no new members are actually created.

Muller and Schwarz (2019) found that higher levels of Twitter exposure have been linked to increased hate crime during Trump’s presidency and aggregate hate crime statistics increased during Trump’s presidency (Beirch, 2019), but these results could be accounted for by changes in hate crime reporting (e.g., greater media attention may have partially corrected hate crime underreporting33). It is important to also emphasize that the causal link between hate group membership and hate crime has not yet been clearly empirically established (Adamczyk et al., 2014; p. 325), though this may simply be an artifact of incomplete data.34

---

33 For a more detailed discussion of issues in hate crime data, see Pezzella, Fetzer, & Keller (2019).

34 Other scholars find that one tactic that is increasingly used by far-right groups is “leaderless resistance,” where groups encourage unaffiliated individuals to engage in ideologically-motivated violence to shield the group and its leader from criminal prosecution (Chermak, Freilich, & Simone, 2010). In other words, it’s possible that a particular violent hate crime could be committed by an individual who closely associates with an
There is, however, a clear opportunity to infer subscription to far-right ideology, as foreshadowed above. The far-right’s reliance on music scenes allows one to study music listenership and track changes in subscription to far-right ideologies. In this paper, I make use of one the largest public music listenership databases (Henning & Reichelt, 2008) to identify top listeners of far-right music. As other scholars found, Trump’s general election victory was associated with increases in posts on neo-Nazi online message boards (e.g., Scrivens et al., 2020), so it logically follows that Trump’s primary victories, xenophobic tweets, and speeches could all be used as opportunities for far-right organizers to influence and recruit acquaintances and friends. Prior research has shown that Last.fm users influence their Last.fm friends’ song choices (Ternovski & Yasseri, 2020); if white supremacists are making use of Trump-related events and far-right music to influence others, we would be able to detect increases in hate music listenership among users who have not previously listened to far-right music.

3.4: Theory

Political elites’ influence on an electorate’s political attitudes does not occur in a vacuum. As with all political communication, other groups and social contacts compete for an individual’s attention (e.g., Zaller 1992). This is true in an online environment as much as it was true when Zaller conducted his studies in 1992. For instance, though former president Donald Trump used Twitter in an unprecedented way for a sitting president (Coe & Park-Ozee 2020), he was one of many voices on social networks. Hate groups also make use of social networks (Gaudette, Scrivens, Venkatesh 2020) and, in some cases, the two actors

organized hate group but has no formal affiliation with the group; in this case, the reported administrative data would fail to link to the hate group to the hate crime.
interacted with one another. For one, there were repeated incidents of Trump retweeting far-right accounts (e.g., Trump retweeted a tweet made by WhiteGenocideTM (Kopan 2016)). And when Trump directly addressed the hate group, Proud Boys, in a television appearance, the hate group quickly interpreted the mention as an endorsement, going so far as to print official Proud Boys t-shirts with Trump’s quote prominently positioned (Palmer 2020).

However, Trump and hate groups do not appear to have the same underlying motivations, which may be why some far-right groups were ultimately “disillusioned” by Trump’s presidency (Einbinder 2019). Trump’s motivation for his rhetoric are to consolidate his own supporters,\(^{35}\) which has tended to include far-right extremists but was ultimately dominated by conservatives with less extreme ideologies. This may explain why Trump did eventually give a speech where he called racism “evil” and hate groups “repugnant” (Merica 2017). Hate groups on the other hand, by and large, have tended to be unwilling to compromise their commitment to racial and xenophobic hatred; as such, membership within such groups has tended to involve a “social stigma from mainstream society” (Jensen, James, & Yates 2020, p. 6).

Despite having motivations that do not necessarily converge, Trump’s statements and behavior stand to benefit far-right groups. As has been extensively documented, exposing Americans to Trump’s rhetoric has changed their perception of anti-prejudice norms, thereby making them increasingly comfortable with expressions of racism and xenophobia (Newman et al. 2021). In other words, the social stigma that members of extremist groups had faced in the past that motivated disengagement and deradicalization

\(^{35}\) As Schneiker (2019) claims, this is accomplished by developing his brand as a “superhero anti-politician celebrity.”
(Jensen, James, & Yates 2020) appears to be weakening. So, under the proposed theoretical framework in this paper, the joint effect of Trump’s normalization of expression of prejudice and hate groups usual recruitment efforts could work in tandem to increase the number of people who actively subscribe to extremist ideologies.

It is important to acknowledge that both political attitudes and prejudices develop from childhood and tend to be very resistant to change (e.g., Sears & Funk 1999; Paluck & Green, 2009). But person-to-person social contact has generally been documented empirically as one of the most effective ways of changing a person’s mind in terms of prejudiced attitudes and behavior (e.g., Munger 2017; Broockman & Kalla 2016; Kalla and Broockman 2020). Far-right groups have used personal social influence for recruitment efforts extensively (e.g., Fording & Cotter 2014) and have since made use of new online spaces for recruitment (e.g., Gaudette, Scrivens, Venkatesh 2020). For instance, qualitative studies found that individuals who became members of far-right groups generally didn’t just “stumble across the material” (Gaudette, Scrivens, Venkatesh 2020, p. 6), but were recruited, usually, by an online friend or social contact who was already a member of some hate group.

As such, I hypothesize that Donald Trump’s erosion of anti-prejudice norms can destigmatize far-right groups, which they, in turn, use as opportunities to recruit others in their broader social network. The main outcome variable of interest in this paper is far-right music listenership, which serves as a proxy for subscription to far-right, white supremacist ideologies. This proxy hinges on the assumption that a person listening to a band like “Racial Purity” actively espouses far-right ideologies that go beyond more common forms of racism.
3.5: Methods and Materials

To evaluate these research questions, I made use of the music social network website, Last.fm, to gather a list of far-right bands, the top listeners of those bands, and the Last.fm friends of those top listeners. I then pulled song plays for all users in the sample. This allowed me to use interrupted time series estimation with Trump-related events as exogenous shocks. In this section, I provide an overview of the data and methods; for greater detail, please see the Supplementary Materials.

First, I used news media sources to compile a list of 22 unanticipated Trump-related events between 2015-2017 (a period when Trump’s many controversies were novel and received intense media coverage). The list of unanticipated events includes 2016 Republican primary victories and widely publicized Twitter controversies (such as him retweeting white supremacist content on Twitter). This list doesn’t include planned events such as the Charlottesville “United the Right” rally. Due to the risk of data dredging, the lists of Trump-related events were pre-registered in the initial Pre-Analysis Plan (PAP) and only those events were used for analysis.

I then collected an initial list of far-right hate bands using Last.fm’s crowdsourced genre tags (e.g., white power, national socialist black metal). To ensure that the music artists were, in fact, far-right, each Last.fm artist page was loaded in a web browser and the genre tags, songs, album titles, the shout-box, the artist biography, pictures of the artists, and related artists were inspected to confirm that every artist was indeed far-right. (The hand-

---

36 Unanticipated Trump-related events are preferred, because far-right groups could recruit both before and after, for instance, a planned Trump rally. I did also compile a list of planned Trump-related events, but, as noted in the original Pre-Analysis Plan, I expected, at best, muted effects.

37 See https://osf.io/z5g27/ or the Supplementary Materials: Original PAP.
coding protocol used can be found in the Supplementary Materials: Data and Materials). The final sample consists of 1,119 hate bands from 43 countries.

I then scraped the top listeners of this set of bands using Last.fm’s API. Midway through the project, Last.fm censored a large subset of artist pages associated with the far-right, so top listeners became inaccessible. And so, an amendment to the PAP was registered.38 As specified in the amended PAP, I made use of archive.org’s “Wayback Machine” to look up historic artist pages and scrape the top listeners of these artists. I then scraped self-reported location and their Last.fm friends. For both the top listeners and their friends, I scraped music listenership histories 2 weeks before and 2 weeks after each of the 22 unanticipated Trump-related events.39

To estimate the impacts of Trump-related events on listenership, I used an interrupted time series (ITS) estimator, which is widely used to measure the impact of some event in time on a running outcome (e.g., see Bernal, Cummins, Gasparrini, 2017; Briesacher et al., 2013). A formalization of the ITS estimator can be found in the Original PAP, but a brief overview follows. To aggregate the data, the time before and after any given Trump-related event is rescaled, such that the cutoff is moved to 0 for every event (c.f., Imbens & Lemieux, 2008). I then regress a running time variable, an indicator that equals 1 for all times after a given Trump-related event (0 otherwise), an interaction of the two, and user and event fixed effects against daily play counts of hate music. For expositional clarity, increases

38 See https://osf.io/z5g27/ or the Supplementary Materials: Amended PAP.

39 I did the same for the list of 10 anticipated Trump-related events in accordance to the Original PAP.
in consumption of far-right music after Trump-related events will be referred to as the “Trump effect” throughout this paper.

This identification strategy rests on the following assumptions. First, I assume that listenership to far-right music does translate to subscription to the far-right ideologies of hate groups. However, an individual may be listening to far-right music, because they like the instrumental parts of the music even though they object to the lyrics. One respondent in a prior qualitative study said that before they espoused far-right ideologies, they listened to far-right music sent to them by friends because they thought the far-right bands were “actually very musically talented. I can pick those things out being a bass player.” (Gaudette, Scrievens, Venkatesh 2020, p. 6). But though, in this example, the listener eventually came to embrace far-right ideologies, other individuals may listen to far-right music without ever joining a far-right group or engaging in any violent behavior. But existing qualitative research suggests that such individuals are uncommon (e.g., Corte & Edwards 2008).

Another key assumption is that the mode of music consumption does not differ before and after Trump-related events. Namely, someone listening to a far-right vinyl record is unlikely to register their listenership on Last.fm, which would imply that that data is missing from our dataset. If individuals went from listening to far-right music on analog media before a Trump event, but listened to far-right music on their computer after, the reported results are overestimated. While this is possible, there is nothing to indicate that this kind of behavior is common among listeners of far-right music.41

40 Though it is theoretically possible to manually add non-digital media listens through Last.fm’s API.

41 I also assume that any observable differences before and after Trump events cannot be explained by false positives. For instance, our list of hate bands includes the prominent white supremacist, punk rock band, Skrewdriver. But Skrewdriver’s first release had no outward connections to racial hate or far-right ideology. It is only in subsequent albums did the band begin to use far-right lyrics and images in their music. There is no
3.6: Results

In this section, I first discuss Last.fm’s data censoring and how my analysis needed to be amended. I then describe my sample. Finally, I preset the main results and a series of placebo checks addressing plausible alternative explanations.

3.6.1: Data Censoring and Amended Analysis

Last.fm unexpectedly started to censor data related to far-right bands, which effectively hid the identity of the top listeners of many far-right bands. As such, I had to additionally collect data from older cached snapshots of the Last.fm website from archive.org. The original PAP relied on scraping data of current top listeners of far-right music; in the amended PAP, I noted that historic top listeners of far-right music would be collected. Upon analyzing the data, I found that a large proportion of historic top listeners of far-right music had ceased to listen to far-right music, which may indicate that they have either disengaged or deradicalized.42,43 This has two key consequences: 1) any analysis of music consumption of far-right group members themselves around Trump-related events severely underpowered, and 2) we have an excellent counterfactual when analyzing the music listenership behavior of friends of these individuals. Namely, I expect only individuals who are active in the far-right music scene to recruit their social contacts on Last.fm. Individuals

---

42 While there is no rigorous estimate of the expected duration of far-right membership on an individual-level, the relatively short half-lives of organized far-right groups (i.e., usually under a year) may be indicative that far-right membership is similarly short-lived (Chermak, Freilich, & Suttmoeller, 2013).

43 Note that these are individuals who have continued to listen to some music on last.fm (hate or non-hate). Individuals who have not listened to any music (hate or non-hate) in our time period of interest are automatically excluded from analysis.
who used to listen to far-right music but have since disengaged are unlikely to try to recruit new members. And yet the unobservable characteristics (such as political preferences) of these two samples are likely similar.\textsuperscript{44} For more details, see Supplementary Materials: Data and Materials, and Amended PAP.\textsuperscript{45}

\textit{3.6.2: Sample}

I was able to recover only 1,368 frequent listeners of hate music (\textquotedblleft top listeners\textquotedblright\textsuperscript{46}) who used their Last.fm account between 2015 and 2017,\textsuperscript{47} and do not reside outside of the United States. For conceptual clarity, I refer to the Last.fm user accounts in the top listener list as \textit{users.} Specifically, a Last.fm account is a user if it has, at one point, appeared in a top listener list of the artist page of any of the far-right bands in our dataset. Users may include:

(1) individuals active in far-right hate groups,

(2) individuals who listen to far-right music without hate group membership,

(3) individuals who have been either in category (1) or (2) but have since reformed and ceased to listen to far-right music,

\textsuperscript{44} Unfortunately, we are unable to confirm this as there was no information about the political preferences or even the demographics of last.fm users (aside from self-reported country of residence).

\textsuperscript{45} While this is mentioned in the amended PAP, the description in the PAP is not very detailed since the number of disengaged far-right listeners was not apparent until after analysis had begun. I did keep lab notes during the data collection and analysis processes, illustrating discussions and the decisions made for analysis, which are available upon request. I also conduct extensive placebo tests, in an effort, to guard against data dredging.

\textsuperscript{46} Last.fm compiles the top listeners of every catalogued music artist and presents a list of \textquotedblleft Top Listener\textquotedblright usernames on the artist\textquoteleft s page.

\textsuperscript{47} This time period does not include all days between 2015-2017—just all days two weeks before and two weeks after each of our unanticipated Trump events.
and (4) individuals who have never listened to far-right music (e.g., they listened to only early albums of a band that had become far-right later in their career).

Since we expect only users active in far-right music scenes will attempt to recruit others, I attempt to limit the analysis to category (1) (or at least categories (1) and (2)). As stated in the amended PAP, there are three key metrics with which to identify users who are most likely to be active in far-right music scenes: 1) how many hate bands they listen to, 2) how often they were classified as top listeners of hate bands, and 3) how central they are in the hate music listeners friends network (as measured by eigenvector centrality).

Users had a total of 45,602 unique friends who do not reside outside of the US. Of those friends, 16,397 were active on Last.fm and have not listened to hate music before; these individuals comprise the primary sample of interest, as these users resemble acquaintances on the periphery that organized far-right groups typically attempt to ideologically influence and recruit (e.g., Woolf & Hulsizer, 2004). For conceptual clarity, I refer to this population as “friends” throughout the main text of this paper.

We expect listenership to change only for friends of users in category (1) and (2) (i.e., users active in the far-right music scene), as users in category (3) and (4) (i.e., users inactive in the far-right music scene) are unlikely to influence and recruit their Last.fm friends. However, there is no precise way to delineate between active and inactive far-right users. To approximate the likelihood that we are looking at the behaviors of friends of users active in far-right music scenes, we can examine how many far-right bands the user listened to, how

---

48 1,185 of these users were friends with other top listeners of hate music, with 2,259 friendships amongst one another.

49 Last.fm friends are analogous to Facebook friends, for a more thorough discussion of friend links, please see Supplementary Materials: Original PAP.
many times they showed up as top listeners of the same hate-band, and how central they are in our network of users. (For a more thorough discussion of the mechanism behind these variables, see Supplementary Materials: Amended PAP). I therefore limit the initial sample of users on these three variables and analyze the listenership behavior of the resulting samples of their friends. As such, I analyze six subgroups of friends of users (see Table 3.1).

Table 3.1: A description of the six samples analyzed

<table>
<thead>
<tr>
<th>Characteristics of Users</th>
<th>Users’ Likelihood of being Active Far-Right Music Listeners</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 (lowest)   2   3   4   5   6 (highest)</td>
</tr>
<tr>
<td>≤ Hate Bands than Mean</td>
<td>≤ Hate Bands than Mean</td>
</tr>
<tr>
<td>≤ Times a Top Listener than Mean</td>
<td>≤ Times a Top Listener than Mean</td>
</tr>
<tr>
<td>≤ Central than Mean</td>
<td></td>
</tr>
</tbody>
</table>

Resulting Sample Size of Friends of Users

| N=9,129 | N=11,436 | N=13,175 | N=3,222 | N=1,442 | N=801 |

3.6.3: Results

In Figure 3.1, I present interrupted time series estimates of hate music listenership after Trump-related events for a series of subgroups conditioning on the above variables. The first three columns depict friends of users who are least likely to be active in far-right music scenes and so they are unlikely to start listening to hate music after Trump-related
events. In column 1, the subgroup only includes friends of users who have been top listeners of hate bands less than average⁵⁰ (twice), have been top listeners of less hate bands than average (two), and are less central in the hate music listenership network than average (eigenvector centrality of 0.07). As expected, for this subgroup, the Trump effect is a precisely estimated zero (95% CI: [-0.002, 0.002]). Columns 2-3 show similar null effects. Since these three columns correspond to friends of likely inactive far-right users, we do not expect a Trump effect. In the proposed theoretical framework, Trump victories alone should not push the average individual to listen to hate music; rather, someone active in the far-right music scene uses the Trump-related events as opportunities to actively influence their social network.

Columns 4-6, in contrast, illustrate Trump effects among friends of likely active far-right users. Specifically, Column 4 restricts the sample to friends of users who appeared as a top listener of more than 2 hate bands (the average) between 2015-2017. Even with this generous restriction and a modest sample size, I recovered a statistically significant (p=0.03) increase in hate music listenership of 0.01 more hate songs per day.⁵¹ Column 5 further restricts the sample to friends of users who were top listeners of the same hate band more than twice (the average). This sample experienced a Trump effect of 0.03 more hate songs per day (p=0.04).⁵² Finally, Column 6 illustrates a sample of friends of users who are most

---

⁵⁰ All reported cutoff averages are calculated such that the unit of analysis is the friend of the hate listener. For robustness checks of these variables, see Supplementary Materials: Robustness Checks.

⁵¹ The interaction of the post-Trump-related-event indicator and the number of hate bands a user’s friend was a top listener of is significant at p<0.001. (To avoid collinearity, I omit user fixed effects in this and subsequent subgroup interaction models.)

⁵² The interaction of the post-Trump-related-event indicator and the number of times a user’s friend was a top listener of the same hate band is not significant (p=0.2), but I believe this variable is imprecise, as it relies on how many times archive.org scraped a Last.fm artist page. Since archive.org scrapes more popular artist pages, less overtly racist bands are likely to be more mainstream and thus garner more archive.org scrapes. As such,
likely to be active in far-right music scenes. For this subgroup, I add a restriction that the user must have a friend who is more central than average in the hate-music-listener network (as measured by eigenvector centrality). The Trump effect for this sample is 0.05 more hate songs per day (p=0.03). As a point of comparison, previous studies found that attending a music concert of a popular (non-hate) artist induces friends of the attendee to listen to 0.06 more songs by that artist immediately after the event (Ternovski & Yasseri, 2020).

Figure 3.1: Friends of top hate listeners who are more active and more central have significant effects

Alternate specifications of this figure can be found the Supplementary Materials (Figures 3.11-3.12).

higher values of this variable do not necessarily translate to more active far-right users. (See Supplementary Materials: Robustness Checks for more details.)

53 The interaction of the post-Trump-related-event indicator and a user’s friend’s eigenvector centrality is significant at the 90% CI (p=0.09). I rounded centrality to the nearest thousandths. The p-value does not meaningfully change with alternative rounding schemes. All users outside of the giant component are treated as zeros.
As expected, when I examine the full sample of friends of users, which includes friends of likely reformed (or otherwise inactive) far-right users, there is no statistically significant Trump effect on hate music listenership (an increase in .002 songs, 95% CI: [-.001, .005]). The large number of users who have not listened to any far-right music in the time period of interest\textsuperscript{54} likely masks the Trump effect.

\textit{3.6.4: Placebo Checks}

Since this is, nevertheless, an observational study, there may be alternative explanations to what I interpret to be a Trump effect. I address several of these alternative explanations through pre-registered placebo checks.

For one, it is not necessarily clear from this analysis that other political events aren’t used in the same way by users active in the far-right music scene. Perhaps members of hate groups use all elections as opportunities to recruit others. As stipulated in the original PAP, I test to see if elections where Trump loses or is not directly involved have the same Trump effects as with unanticipated Trump-related events. As such, I compiled a list of all state and federal election events in the time period of analysis that did not coincide with either the unanticipated or the planned Trump-related events.\textsuperscript{55} Because these “placebo” events are still temporally proximal to Trump-related events, we are more likely to see directionally positive differences (i.e., due to lingering Trump effects). Instead, as seen in Figure 3.2, there were directionally negative effects for friends of users most likely active in far-right music scenes. This may be due to the fact that the placebo events include Trump losses.

\textsuperscript{54} Only 52\% of all users have listened to \textit{any} hate music in the time period of interest (576 days in total).

\textsuperscript{55} See the Supplementary Materials for a full list of the Placebo Events.
Figure 3.2: Impact of placebo events on hate music listenership of friends of far-right users

As stipulated in my original PAP, another registered placebo test looks at the impact of Trump-related events on friends of frequent hate music listeners in other countries. Though some scholars have found that Trump’s general election victory led to increases in racist attitudes in Europe (Giani & Meon, 2019), I believe that such effects are unlikely to be so granular as to reflect primary election wins and controversial tweets. As seen in Figure 3.3, there are indeed null effects for all subgroups.\(^{56}\)

\(^{56}\) Since Giani & Meon (2019) found that Trump’s rise may influence racial attitudes in Europe and many of the hate bands in my sample are from Europe, as a robustness check, I also limit this placebo test to residents of South American countries where it is unlikely that far-right music listeners would choose Trump-related events as opportunities to recruit others, I obtain similarly null effects for all subgroups. That said, it is worth noting that even Latin American countries appears to have active far-right music scenes. In coding far-right bands from Mexico, I found that they similarly use overt Nazi iconography (e.g., swastikas), their song lyrics are rife
Figure 3.3: Impact of Trump-related events on hate music listenership of non-US friends of far-right users

Finally, I examined whether Trump-related events had any impacts on non-hate music consumption (i.e., all music artists who were not on our list of far-right bands). As seen in Figure 3.4, friends of likely active far-right users did not see similar increases in non-hate music listenership after Trump-related events. It is worth noting that Trump-related activities with antisemitism and racial hatred, and they have released collaborations with European white supremacist groups (e.g., “Aryan Roots (Mexican-Belgian Axis”).

57 The original PAP stipulated that we use mainstream (i.e., Top 40) and similar niche music as placebos to test the hypothesis that users do not modify their overall (i.e., non-hate) music consumption as a result of Trump events, but this decision was due to Last.fm’s API limitations. The API initially allowed for look-ups based on user name and artist name. Midway through analysis, Last.fm eliminated that functionality, so I had to scrape the entirety of a user’s listenership history in the time windows of interest. As such, I have recovered a placebo that better reflects overall music consumption (i.e., all non-hate music that a user listened to) and this is the placebo dependent variable I use the analyses reported here.
events appeared to reduce consumption of music among users who are friends with users who are likely inactive in the far-right scene.

Figure 3.4: Impact of Trump-related events on non-hate music listenership of friends of far-right users

Additional analyses, including robustness checks and placebo tests can be found in the Supplementary Materials.

3.7: Discussion and Conclusions

The empirical results are consistent with my proposed theoretical framework: that likely members of far-right groups did use Trump’s primary victories and controversies to influence and recruit their friends on Last.fm. However, these results have notable limitations. The biggest drawback is that I have no way of documenting the communication
between far-right users and their Last.fm friends. It is entirely possible that friends of active far-right users may have already had far-right political preferences and a public Trump event was all that was necessary to embolden them to listen to hate music quasi-publicly. However, what makes this explanation less likely is that friends of inactive far-right users do not show similar increases in hate music listenership.58

It is also important to emphasize that hate music listenership does not necessarily translate to hate group membership or even an espousal of far-right ideology. Individuals may simply listen to far-right music without endorsing its message. Currently, there are no rigorous studies estimating how many individuals listen to far-right music without embracing the expressed ideologies. Regardless, given that many of the bands often have explicit lyrics with the goal of inciting racial hatred, increases in consumption of this material are nevertheless alarming.

Still, despite these limitations, this study adds another perspective to the literature on how elites and organizations can affect political attitudes of the electorate. This paper speaks to the extremes—namely, how far-right hate groups can exacerbate the social impacts of a president who showed little restraint in communicating extremist content. Future research should better establish my proposed causal pathway; I assert that when elites erode anti-prejudice norms, extremist hate groups use the opportunity to recruit others in their social networks. If the reader is unconvinced by the causal identification strategy in this paper, I argue that this paper, at the very least, illustrates that friends of active listeners of far-right music themselves listen to far-right music at higher rates after unanticipated Trump-related

58 While this placebo test makes this explanation less likely, my results may still be explained by homophily. Namely, individuals who are friends with active far-right users may simply be systematically different (i.e., more far-right) from individuals who are friends with inactive far-right users.
events. So even if the results in this paper are not evidence of active recruitment, they further support the theory that Trump didn’t just erode norms, but emboldened individuals to listen to music that incites racial and xenophobic hatred and violence.

Finally, this paper illustrates the usefulness of an alternative behavioral data source to estimate subscription to far-right ideologies and that such data can be used on a large scale. Given the difficulty of measuring far-right group activity and the proliferation of data of online behavior, it may be possible to estimate far-right group membership using similar trace data on other platforms (such as YouTube) without needing to parse and interpret the content of, for instance, video blogs.

3.8: References


Southern Law Poverty Center. Retrieved from


https://www.theguardian.com/world/2014/sep/10/azov-far-right-fighters-ukraine-neo-nazis


3.9: Supplementary Materials

3.9.1: Data and Materials

3.9.1.1: Trump-Related Events

A list of Trump-related events was compiled using media sources and registered with our original PAP. There were two types of events: 1) unanticipated events and 2) planned events. Unanticipated events are unexpected victories, speeches, tweets, or announcements. Planned events are events where the timing is known in advance (such as rallies).

3.9.1.1.1: Full List of Unanticipated Trump-Related Events

1. February 28, 2015 – CNN Interview in an answer about David Duke and the KKK: “Well, just so you understand, I don’t know anything about David Duke, OK? I don’t know anything about what you’re even talking about with white supremacy or white supremacists. So I don’t know.” (Trump, 2015 qtd. in Finnegan & Barabak, 2018)

2. June 16, 2015 – Tweet: “When Mexico sends its people, they’re not sending their best. They’re not sending you. They’re not sending you. They’re sending people that have lots of problems….They’re bringing drugs. They’re bringing crime. They’re rapists, and some, I assume, are good people.” (Trump, 2015 qtd. in Finnegan & Barabak, 2018)


5. February 20, 2016 – Won South Carolina Primary, “clears path to nomination”
   (Jacobs, Bixby, Siddiqui, Sullivan & Gabbatt, 2016)

6. February 23, 2016 – Won Nevada Caucus, “further solidifies standing as the front-
   runner for the GOP nomination” (Goldmacher, 2016)

7. February 28, 2016 – “Donald Trump Retweets Post with Quote from Mussolini”
   (Haberman, 2016)

8. March 1, 2016 – Super Tuesday, “With a big Super Tuesday, Trump has the
   Republican nomination in his sights” (Barabak, 2016)

9. March 15, 2016 – Trump wins Florida, Rubio drops out. (Mazzei, Sherman, & Clark,
   2016)

10. May 3, 2016 – Ted Cruz drops out of race leaving “Donald Trump as the only
    candidate capable of clinching the nomination outright” (Glueck & Goldmacher,
    2016)

11. May 26, 2016 – “Trump reaches delegate count needed to clinch Republican
    nomination” (McCarthy, 2016)

    of a club or society very strongly pro-Mexican” (Trump, 2016 qtd. in Finnegan &
    Barabak, 2018)

13. July 2, 2016 – Trump retweets anti-Clinton graphic previously posted “on an anti-
    Semitic, white supremacist message board” (Diamond, 2016).

14. July 5, 2016 – Trump tweets about “setting the record for the most GOP primary
    votes ever” (Doran, 2016).

15. October 7, 2016 – Publication of recording where “Trump brags about groping
    women” (Graham, 2016).


21. November 29, 2017 – Trump retweet of far-right Nationalist who was “recently arrested for inciting hatred and violence against Muslims.” (Finnegan & Barabak, 2018)

22. December 23, 2017 – New York Times report claims Trump said Haitians “all have AIDS” and Nigerians “[o]nce they had seen the United States… would never `go back to their huts’ in Africa” (Shear & Davis, 2017; Finnegan & Barabak, 2018)

3.9.1.2: Full List of Planned Trump-Related Events

1. December 7, 2015 – South Carolina rally, calling for “total and complete shutdown of Muslims entering the United States until our country’s representatives can figure out what the hell is going on.” (Trump, 2015 qtd. in Finnegan & Barabak, 2018)


5. August 9, 2016 – North Carolina rally, Trump “suggesting violence against Mrs. Clinton or liberal jurists” (Corasaniti & Haberman, 2016)

6. September 26, 2016 – Second Presidential debate, Trump says that if he were in charge of the law Clinton “would be in jail” (Roberts, Jacobs, & Siddiqui, 2016)


8. August 15, 2017 – Press conference, Trump on the rioting from the pro-Confederate rally in Charlottesville: “I think there is blame on both sides….You also had people that were very fine people on both sides….Not all of those people were neo-Nazis, believe me. Not all of those people were white supremacists by any stretch.” (Trump, 2017 qtd. in Finnegan & Barabak, 2018)

9. August 22, 2017 – Phoenix rally, Trump on the removal of Confederate monuments: “They’re trying to take away our culture. They’re trying to take away our history. And our weak leaders, they do it overnight. These things have been there for 150 years, for a hundred years. You go back to a university and it’s gone. Weak, weak people.” (Trump, 2017 qtd. in Finnegan & Barabak, 2018)

10. September 22, 2017 – Alabama rally, Trump on the black football players protesting racial discrimination during National Anthem: “Wouldn’t you love to see one of these NFL owners, when somebody disrespects our flag, to say, ‘Get that son of a bitch off the field right now. Out. He’s fired. He’s fired!’” (Trump, 2017 qtd. in Finnegan & Barabak, 2018)
3.9.1.2: Hate Music Artists

We first collected a sample of far-right artists on Last.fm by querying popular hate music genres\footnote{\textit{Last.fm} uses a crowdsourced genre tagging system, where every registered Last.fm user is able to tag an artist as being in a particular music genre. The most popular tags become the primary music genres for that artist.} on Last.fm’s API. The genres included are rac, white power, hatecore, rock against communism, nsbm, or ns black metal.\footnote{There were other hate music genres, but they had high levels of vandalism (i.e., mistagged artists).} This yielded approximately 6,097 unique bands.

Midway through our research, Last.fm had taken action against hate music pages and censored a large subset of artist pages associated with the far-right. In Figure 3.5, we can see the difference between a censored page and an uncensored artist page. Furthermore, in Figure 3.6, we can see the effect of this censorship policy on the Listeners pages of both bands.\footnote{These changes are also reflected in Last.fm’s API.}

As such, we made use of Last.fm’s lengthy history online and archive.org’s Internet archiving project to compile a list of top listeners of hate music. Using the list of hate bands that we had compiled before Last.fm’s censorship policy went into effect, we extract all listeners associated with those hate bands on archive.org. (We draw on \textit{all} relevant Last.fm snapshots that are available on archive.org.) We then narrow our sample of hate music artists to only those artists for whom we were able to scrape top and/or recent listeners; this yields 2,018 artists.
A closer look at our list of hate bands revealed that despite the precautions we took to avoid vandalized genre tags, there were artists who were clearly mistagged (e.g., j-pop artist, Kahimi Karie) or spurious (e.g., Geico [the Insurance Company]). Additionally, we noted that some hate bands had the same name as artists who had no connection to far right ideologies (e.g., Evil). As such, for every artist where we were able to successfully extract
listeners from archive.org, we manually looked up the Last.fm artist page in a web browser and inspected the genre tags, songs, album titles, the shout-box, the artist biography, pictures of the artists, and related artists. We then hand-coded them as either “mistagged/fake”, “suspect”, or “multi.” Mistagged/fake was used in cases where it was clear that the artist was either not a hate music artist (e.g., several Eurovision pop stars) or not a real music artist (e.g., politicians, cartoon characters, etc.). “Suspect” was used in cases where the connection to far-right ideologies was either disputed (based on the shout-box activity and/or artist biography), the band was a “troll” artist (e.g., certain grindcore bands), or the ideologies were not immediately clear (e.g., anti-racist hatecore). Finally, an artist was tagged “multi” if the biography or the shout-box made it clear that multiple artists shared the same band name.62 To reduce the amount of noise in our data, we exclude all artists (and the corresponding listeners) who were tagged as either “mistagged/fake,” “suspect,” or “multi.” This yields a total of 1,119 unique artists.

3.9.1.2.1: Hand-Coding Protocol Used

1. Go to Last.fm page for each band
2. Fill in first column with “sole” “multi” “fake” or “mistagged”
3. Fill in second column with “suspect” or leave blank

Coding protocol:

IF censored (no images, no comments, no top listeners)

MARK: “sole”

---

62 There were a few cases where multiple artists shared the same name, but all were explicitly far-right (e.g., Ahnenerbe, originally the name of a Nazi pseudoscientific institute that was to research the supposed archaeological and cultural history of the hypothesized "Aryan race"); these cases were not tagged as “multi,” since, for our purposes, they are no different from uniquely named far-right bands.
IF bio indicates there’s only one band

MARK: “sole”

IF no bio, but no clear evidence that there are more than one band

MARK: “sole”

IF bio, pictures or shout box activity indicates that there are multiple bands with the same name

MARK: “multi”

IF page indicates that the band is not a real music artist (e.g. novelty Youtube videos, advertisements, historical figures who never made music, celebrities that never made music)

MARK: “fake”

IF tags, shout box activity, photos, album titles and song titles CLEARLY demonstrate that the artist has no ties to far-right music. There should be not a single doubt here.

MARK: “mistagged”

IF there is no evidence of far-right ideologies OR nazi status disputed

MARK: “suspect”

KEY WORDS TO LOOK FOR: “NSBM, National Socialist, white nationalist, white pride, white power, RAC, rock against communism, hatecore, skinhead, oi, racist, nazi, 14/88, 88, 14”, photos with censored faces, usually all white males (though there are some exceptions)
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>GERMANY</td>
<td>221</td>
<td>19.75</td>
<td>19.75</td>
</tr>
<tr>
<td>US</td>
<td>126</td>
<td>11.26</td>
<td>31.01</td>
</tr>
<tr>
<td>RUSSIA</td>
<td>122</td>
<td>10.9</td>
<td>41.91</td>
</tr>
<tr>
<td>POLAND</td>
<td>105</td>
<td>9.38</td>
<td>51.3</td>
</tr>
<tr>
<td>FRANCE</td>
<td>47</td>
<td>4.2</td>
<td>55.5</td>
</tr>
<tr>
<td>SWEDEN</td>
<td>42</td>
<td>3.75</td>
<td>59.25</td>
</tr>
<tr>
<td>UKRAINE</td>
<td>41</td>
<td>3.66</td>
<td>62.91</td>
</tr>
<tr>
<td>UK</td>
<td>40</td>
<td>3.57</td>
<td>66.49</td>
</tr>
<tr>
<td>ITALY</td>
<td>35</td>
<td>3.13</td>
<td>69.62</td>
</tr>
<tr>
<td>CANADA</td>
<td>30</td>
<td>2.68</td>
<td>72.3</td>
</tr>
<tr>
<td>HUNGARY</td>
<td>30</td>
<td>2.68</td>
<td>74.98</td>
</tr>
<tr>
<td>SERBIA</td>
<td>24</td>
<td>2.14</td>
<td>77.12</td>
</tr>
<tr>
<td>FINLAND</td>
<td>23</td>
<td>2.06</td>
<td>79.18</td>
</tr>
<tr>
<td>CZECHIA</td>
<td>22</td>
<td>1.97</td>
<td>81.14</td>
</tr>
<tr>
<td>GREECE</td>
<td>18</td>
<td>1.61</td>
<td>82.75</td>
</tr>
<tr>
<td>SLOVAKIA</td>
<td>18</td>
<td>1.61</td>
<td>84.36</td>
</tr>
<tr>
<td>BELARUS</td>
<td>15</td>
<td>1.34</td>
<td>85.7</td>
</tr>
<tr>
<td>BELGIUM</td>
<td>15</td>
<td>1.34</td>
<td>87.04</td>
</tr>
<tr>
<td>SPAIN</td>
<td>15</td>
<td>1.34</td>
<td>88.38</td>
</tr>
<tr>
<td>AUSTRALIA</td>
<td>13</td>
<td>1.16</td>
<td>89.54</td>
</tr>
<tr>
<td>BULGARIA</td>
<td>12</td>
<td>1.07</td>
<td>90.62</td>
</tr>
<tr>
<td>CROATIA</td>
<td>11</td>
<td>0.98</td>
<td>91.6</td>
</tr>
<tr>
<td>MEXICO</td>
<td>11</td>
<td>0.98</td>
<td>92.58</td>
</tr>
<tr>
<td>MULTI</td>
<td>11</td>
<td>0.98</td>
<td>93.57</td>
</tr>
<tr>
<td>BRAZIL</td>
<td>10</td>
<td>0.89</td>
<td>94.46</td>
</tr>
<tr>
<td>NETHERLANDS</td>
<td>9</td>
<td>0.8</td>
<td>95.26</td>
</tr>
<tr>
<td>NORWAY</td>
<td>8</td>
<td>0.71</td>
<td>95.98</td>
</tr>
<tr>
<td>ARGENTINA</td>
<td>7</td>
<td>0.63</td>
<td>96.6</td>
</tr>
<tr>
<td>JAPAN</td>
<td>5</td>
<td>0.45</td>
<td>97.05</td>
</tr>
<tr>
<td>OTHER</td>
<td>5</td>
<td>0.45</td>
<td>97.5</td>
</tr>
<tr>
<td>AUSTRIA</td>
<td>4</td>
<td>0.36</td>
<td>97.86</td>
</tr>
<tr>
<td>ESTONIA</td>
<td>4</td>
<td>0.36</td>
<td>98.21</td>
</tr>
<tr>
<td>PORTUGAL</td>
<td>4</td>
<td>0.36</td>
<td>98.57</td>
</tr>
<tr>
<td>SWITZERLAND</td>
<td>4</td>
<td>0.36</td>
<td>98.93</td>
</tr>
<tr>
<td>CHILE</td>
<td>3</td>
<td>0.27</td>
<td>99.2</td>
</tr>
<tr>
<td>PERU</td>
<td>3</td>
<td>0.27</td>
<td>99.46</td>
</tr>
<tr>
<td>BOSNIA</td>
<td>2</td>
<td>0.18</td>
<td>99.64</td>
</tr>
<tr>
<td>IRELAND</td>
<td>2</td>
<td>0.18</td>
<td>99.82</td>
</tr>
<tr>
<td>SOUTH AFRICA</td>
<td>2</td>
<td>0.18</td>
<td>100</td>
</tr>
<tr>
<td>Total</td>
<td>1,119</td>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>
3.9.1.3: Top Listeners of Hate Music

Ideally, we would want data from recent top listeners of hate music, but we were only able to extract historical top listenership data from archive.org. Though we found that 85% of the accounts we scraped from archive.org still existed at the time of analysis (16,396 out of an initial 19,236), once we exclude all non-US nationals, a majority of the initial archive.org sample appeared to be lapsed users. Only 28% of the 5,038 users in our sample of interest have listened to any music (far-right or otherwise) in our time windows of interest. A summary of these exclusions is found in the Table 3.3.

Table 3.3: Top listeners of hate music after exclusions

<table>
<thead>
<tr>
<th></th>
<th>Hate Music Listeners</th>
</tr>
</thead>
<tbody>
<tr>
<td>Archive.org Total</td>
<td>19,236</td>
</tr>
<tr>
<td>After Excluding Banned/Deleted Accounts</td>
<td>16,396</td>
</tr>
<tr>
<td>After Excluding Non-US Nationals</td>
<td>5,038</td>
</tr>
<tr>
<td>After Excluding Lapsed Accounts</td>
<td>1,368</td>
</tr>
</tbody>
</table>

3.9.1.4: Friends of Top Listeners of Hate Music

We pulled the complete friends list of each of the 1,368 hate music listeners in our sample. We found that only 94 hate music listeners did not have Last.fm friends. This yields a network of 119,577 friends of hate music listeners. The average hate music listener in our sample has an average of 87.4 friends; the median hate music listener has 28 friends. Of those 119,577 friends, 89,621 friends are unique (i.e., some hate music listeners have the same friends). As illustrated in Table 3.4, we then exclude all subjects who affirmatively self-
reported living outside of the United States,\textsuperscript{63} which leaves 45,602 participants. Of those participants, 34,719 are not themselves top listeners of hate music. After excluding users who have not listened to any music during the time period of interest, we are left with 16,756 unique active users. An additional 313 of those subjects listened to hate music previously but were not classified as top listeners.

Table 3.4: Friends of top listeners of hate music after exclusions

<table>
<thead>
<tr>
<th>Friends of Hate Music Listeners</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Unique Friends Total</td>
<td>89,621</td>
</tr>
<tr>
<td>After Excluding Non-US Nationals</td>
<td>45,602</td>
</tr>
<tr>
<td>After Excluding Friends Who Themselves are Top Hate Music Listeners</td>
<td>34,719</td>
</tr>
<tr>
<td>After Excluding Friends Who are Lapsed Listeners</td>
<td>16,756</td>
</tr>
<tr>
<td>After Excluding Friends Who Listened to Hate Music Previously</td>
<td>16,443</td>
</tr>
</tbody>
</table>

3.9.2: Issues with Naïve Sample of Top Listeners of Hate Music

There is evidence to indicate that the 1,368 sample of “top” listeners of hate music includes many users who are not current, frequent listeners of hate music. This sample only devoted 2% of their daily music consumption to hate music on average. As seen in Figure OA3, the decreasing trend may indicate survivorship bias: only users who are ideologically motivated persisted in our sample, whereas other users ceased to listen to far-right music—perhaps they abandoned far-right views or never held them in the first place (i.e., some far-right artists have releases that do not qualify as hate music). Most importantly, Last.fm does not specify how much music a user must listen to appear as a “top listener” for a particular

\textsuperscript{63} As stipulated in our original PAP, we have a loose definition of likely-US nationals. They include users who affirmatively declared that they are from the US, but also users whose country data is missing and users who specify likely spurious location data (e.g., Antarctica, DPRK, Vatican City).
artist. For obscure hate music bands, it may have been possible to listen to the artist just a few times and end up in our top listener sample. We find that only 52% of all users classified as top listeners of hate music have listened to any hate music in the time period of interest (576 days in total).

**Figure 3.7**: Percent of daily listening that is hate music in our naïve sample of top listeners of hate music

![Graph showing the percent of daily listening that is hate music from 01 Jan 2015 to 01 Jan 2017.](image)

*Each point denotes the daily average across all users in the sample. The horizontal line represents the average across all days.*

The distribution of hate music listenership is highly unequal. The top 1% of hate music listeners have listened to over 50% of all hate music listened. Hate music consumption in this sample has a Gini coefficient of .87, whereas non-hate music has a Gini coefficient of .68. As specified in our amended PAP, there are several ways we can reduce attenuation bias in our data to the population of active far-right listeners (without conditioning on a post-treatment outcome). Namely, we can limit our analysis to only those individuals who have appeared as top listeners:
1) of more hate bands than average

2) of the same hate band more times than average.

As seen from Figures 3.8-3.9, both approaches produce samples with lower levels of attenuation bias. As seen in Figure 3.10, combining the two criteria produces the best sample of high-frequency hate music listeners, with nearly 10% of users’ daily listening devoted to hate music.

**Figure 3.8**: Percent of daily listening that is hate music among users who were top listeners of more hate bands than average

*Each point denotes the daily average across all users in the sample. The horizontal line represents the average across all days.*
Figure 3.9: Percent of daily listening that is hate music among users who were top listeners of the same hate band more times than average.

Each point denotes the daily average across all users in the sample. The horizontal line represents the average across all days.
Figure 3.10: Percent of daily listening that is hate music among users who were top listeners of the same hate band more than average and more hate bands than average.

Each point denotes the daily average across all users in the sample. The horizontal line represents the average across all days.

3.9.3: Registered Analyses in Original PAP

Our original PAP stated that we would analyze the impact of Trump-related unanticipated events, planned events, and the combination of both on hate music listenership (“Trump effects”). However, as noted in the original PAP, any Trump effects due to planned events are likely to be dampened by users anticipating the events. The change in Last.fm policy regarding hate bands left us severely underpowered, so an effect that we expect is already-muted also has greater imprecision from reduced sample size. As such, it is not surprising that the impact of planned events and the combination of the unanticipated and planned events are not statistically significant. Due to the attenuation bias in our sample of top listeners of far-right music, even unanticipated events only approach significance. But before completing the collection of this data, we did register an amended PAP, which is
what we abide by for our main analyses (reported in the main paper). For completeness, we nevertheless report the analyses specified in our original PAP below.

**Table 3.5:** Trump effects of planned events

<table>
<thead>
<tr>
<th></th>
<th>H1a</th>
<th>H1b</th>
<th>H1c</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Time</strong></td>
<td>0.0029</td>
<td>0.0002</td>
<td>0.0002</td>
</tr>
<tr>
<td></td>
<td>(0.0020)</td>
<td>(0.0002)</td>
<td>(0.0001)</td>
</tr>
<tr>
<td><strong>Post-event indicator</strong></td>
<td>0.0397</td>
<td>-0.0006</td>
<td>-0.0019</td>
</tr>
<tr>
<td></td>
<td>(0.0258)</td>
<td>(0.0019)</td>
<td>(0.0016)</td>
</tr>
<tr>
<td><strong>Time*Post-event indicator</strong></td>
<td>-0.0050</td>
<td>0.0001</td>
<td>0.0003</td>
</tr>
<tr>
<td></td>
<td>(0.0032)</td>
<td>(0.0002)</td>
<td>(0.0002)</td>
</tr>
<tr>
<td><strong>N1 (User-Event-Days)</strong></td>
<td>383,040</td>
<td>3,047,240</td>
<td>2,982,980</td>
</tr>
<tr>
<td><strong>N2 (Users)</strong></td>
<td>1,368</td>
<td>14,442</td>
<td>14,162</td>
</tr>
</tbody>
</table>

H1a describes effects among our “Top Listener” sample. H1b describes effects among friends of our “Top Listener” sample who are themselves not in the “Top Listener” sample. H1c describes effects among friends of our “Top Listener” sample who have not listened to hate music previously.

**Table 3.6:** Trump effects of unanticipated events

<table>
<thead>
<tr>
<th></th>
<th>H1a</th>
<th>H1b</th>
<th>H1c</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Time</strong></td>
<td>0.0033</td>
<td>-0.0001</td>
<td>-0.0002</td>
</tr>
<tr>
<td></td>
<td>(0.0015)</td>
<td>(0.0001)</td>
<td>(0.0001)</td>
</tr>
<tr>
<td><strong>Post-event indicator</strong></td>
<td>-0.0289</td>
<td>0.0008</td>
<td>0.0020</td>
</tr>
<tr>
<td></td>
<td>(0.0190)</td>
<td>(0.0016)</td>
<td>(0.0014)</td>
</tr>
<tr>
<td><strong>Time*Post-event indicator</strong></td>
<td>-0.0007</td>
<td>0.0000</td>
<td>0.0001</td>
</tr>
<tr>
<td></td>
<td>(0.0024)</td>
<td>(0.0002)</td>
<td>(0.0002)</td>
</tr>
<tr>
<td><strong>N1 (User-Event-Days)</strong></td>
<td>842,688</td>
<td>6,866,160</td>
<td>6,718,824</td>
</tr>
<tr>
<td><strong>N2 (Users)</strong></td>
<td>1,368</td>
<td>16,710</td>
<td>16,397</td>
</tr>
</tbody>
</table>

H1a describes effects among our “Top Listener” sample. H1b describes effects among friends of our “Top Listener” sample who are themselves not in the “Top Listener” sample. H1c describes effects among friends of our “Top Listener” sample who have not listened to hate music previously.
Table 3.7: Trump effects of unanticipated and planned events

<table>
<thead>
<tr>
<th></th>
<th>$H,1a$</th>
<th>$H,1b$</th>
<th>$H,1c$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Time</strong></td>
<td>0.0032</td>
<td>0.0000</td>
<td>-0.0001</td>
</tr>
<tr>
<td></td>
<td>(0.0012)</td>
<td>(0.0001)</td>
<td>(0.0001)</td>
</tr>
<tr>
<td><strong>Post-event Indicator</strong></td>
<td>-0.0075</td>
<td>0.0004</td>
<td>0.0008</td>
</tr>
<tr>
<td></td>
<td>(0.0154)</td>
<td>(0.0013)</td>
<td>(0.0011)</td>
</tr>
<tr>
<td><strong>Time*Post-event Indicator</strong></td>
<td>-0.0021</td>
<td>0.0000</td>
<td>0.0001</td>
</tr>
<tr>
<td></td>
<td>(0.0019)</td>
<td>(0.0002)</td>
<td>(0.0001)</td>
</tr>
<tr>
<td><strong>N1 (User-Event-Days)</strong></td>
<td>1,225,728</td>
<td>9,913,400</td>
<td>9,701,804</td>
</tr>
<tr>
<td><strong>N2 (Users)</strong></td>
<td>1,368</td>
<td>16,756</td>
<td>16,443</td>
</tr>
</tbody>
</table>

$H\,1a$ describes effects among our “Top Listener” sample. $H\,1b$ describes effects among friends of our “Top Listener” sample who are themselves not in the “Top Listener” sample. $H\,1c$ describes effects among friends of our “Top Listener” sample who have not listened to hate music previously.

Table 3.8: Anticipation effects for planned events among top listeners of hate music

<table>
<thead>
<tr>
<th></th>
<th>No Lead</th>
<th>1-Day Lead</th>
<th>2-Day Lead</th>
<th>3-Day Lead</th>
<th>4-Day Lead</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Time</strong></td>
<td>0.0029</td>
<td>0.0027</td>
<td>0.0022</td>
<td>0.0015</td>
<td>0.0013</td>
</tr>
<tr>
<td></td>
<td>(0.0020)</td>
<td>(0.0022)</td>
<td>(0.0025)</td>
<td>(0.0028)</td>
<td>(0.0032)</td>
</tr>
<tr>
<td><strong>Post-event Indicator</strong></td>
<td>0.0397</td>
<td>0.0344</td>
<td>0.0341</td>
<td>0.0360</td>
<td>0.0328</td>
</tr>
<tr>
<td></td>
<td>(0.0258)</td>
<td>(0.0256)</td>
<td>(0.0266)</td>
<td>(0.0290)</td>
<td>(0.0328)</td>
</tr>
<tr>
<td><strong>Time*Post-event Indicator</strong></td>
<td>-0.0050</td>
<td>-0.0038</td>
<td>-0.0027</td>
<td>-0.0015</td>
<td>-0.0005</td>
</tr>
<tr>
<td></td>
<td>(0.0032)</td>
<td>(0.0032)</td>
<td>(0.0032)</td>
<td>(0.0034)</td>
<td>(0.0036)</td>
</tr>
<tr>
<td><strong>N1 (User-Event-Days)</strong></td>
<td>383,040</td>
<td>383,040</td>
<td>383,040</td>
<td>383,040</td>
<td>383,040</td>
</tr>
<tr>
<td><strong>N2 (Users)</strong></td>
<td>1,368</td>
<td>1,368</td>
<td>1,368</td>
<td>1,368</td>
<td>1,368</td>
</tr>
</tbody>
</table>

3.9.4: Registered Analyses in Amended PAP

The results of the amended PAP are presented in the paper. The below figures illustrate alternative sample specifications.

---

64 We do not pursue the anticipation effect analysis with unanticipated events, since the effect of unanticipated Trump-related events for our “Top Listeners” sample was not statistically significant, so the procedure specified in our original PAP is not applicable.
Figure 3.11: Friends of top hate listeners who are more active and more central have statistically significant Trump effects (Alternate 1)
Figure 3.12: Friends of top hate listeners who are more active and more central have statistically significant Trump effects (Alternate 2)
3.9.5: Robustness Checks

3.9.5.1: Attenuation of the Trump Effect on Friends of Top Listeners of Hate Music

In our amended PAP, we used three variables to identify active and (potentially) influential members of the far-right music scenes: 1) the number of hate bands the user was a top listener of, 2) the number of times the user was a top listener of the same hate band, and 3) the eigenvector centrality of the far-right user in the Last.fm friends network of top listeners of hate music.

As noted in the main text, there is reason to believe that the number of times a user was a top listener of the same hate band may be an imperfect measure of how embedded a user is in a far-right music scene. From Figure 3.14, we see that this may indeed be the case: higher levels of listenership of the same hate band does not seem to translate to larger peer-influence effects. This is in stark contrast to the other two variables illustrated in Figure 3.13 and 3.15, which illustrate clearly increasing trends. It is not obvious why this variable is a poor predictor of social influence. Some possible explanations include: these users may be fans of non-racist material from some of the more popular hate artists (e.g., Skrewdriver’s first album) or they may listen to far-right music without embracing the far-right ideology.
Figure 3.13: Attenuation of Trump effect on friends of far-right listeners across number of hate bands the far-right user was a top listener of.

Figure 3.14: Attenuation of Trump effect on friends of far-right listeners across number of times the far-right user was a top listener of the same hate band.
**Figure 3.15**: Attenuation of Trump Effect on friends of far-right listeners across far-right user’s eigenvector centrality in hate music listener network

![Figure 3.15: Attenuation of Trump Effect on friends of far-right listeners across far-right user’s eigenvector centrality in hate music listener network](image)

### 3.9.5.2: *Alternative Variables to Address Attenuation*

There are variables that may be used to identify Last.fm users active in far-right music scenes besides the ones registered in our amended PAP. These include 1) the most recent date a user was a top listener of a hate band, and 2) the degree of the far-right user in the top listener Last.fm friends network (i.e., a user’s number of friends who are also top listeners of hate music). We run two interaction models to determine if these alternative variables can reduce attenuation bias in our sample by screening for only active hate music users. Though recency does not predict higher peer influence effects (p=.63), degree does (p=.049).

### 3.9.5.3: *Alternative Dependent Variable*

To determine if users are increasing their total music listenership to include hate bands or are substituting non-hate-bands for hate-bands, we check our main effects using a different dependent variable: the percent of daily tracks listened that are hate music. As seen
in the Figure 3.16, we find suggestive evidence that our results are driven by a substitution effect.

**Figure 3.16:** Trump effect on friends of far-right listeners in terms of % of daily tracks that are hate music.
3.9.6: Placebo Checks

This section covers three types of placebos: 1) music, 2) events, and 3) users.

3.9.6.1: Music Placebo

Our original PAP stipulated that we use mainstream (i.e., Top 40) and similar niche music as placebos to test the hypothesis that users do not modify their overall (i.e., non-hate) music consumption as a result of Trump events, but this decision was due to Last.fm’s API limitations. The API initially allowed for look-ups based on user name and artist name. Midway through analysis, Last.fm eliminated that functionality, so we had to scrape the entirety of a user’s listenership history in our time windows of interest. As such, we have a placebo that better reflects overall music consumption (i.e., all non-hate music that a user listened to) and this is the placebo dependent variable we use in all subsequent analyses reported here.
Figure 3.17: Impact of Trump-related events on non-hate music listenership of friends of far-right users.

It is worth noting that Trump-related events appeared to reduce consumption of music among users who are friends with users who are likely inactive in the far-right scene.
For completeness and to abide by our original PAP, we also include placebo tests that correspond to Hypotheses 3-4 below.

**Table 3.9:** Impact of planned events on non-hate music listenership

<table>
<thead>
<tr>
<th></th>
<th>$H_{1a}$</th>
<th>$H_{1b}$</th>
<th>$H_{1c}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Time^{\text{Post-event Indicator}}$</td>
<td>0.0175</td>
<td>0.0077</td>
<td>0.0061</td>
</tr>
<tr>
<td>($0.0167$)</td>
<td>($0.0086$)</td>
<td>($0.0086$)</td>
<td></td>
</tr>
<tr>
<td>$Post-event Indicator$</td>
<td>0.2842</td>
<td>0.0597</td>
<td>0.0416</td>
</tr>
<tr>
<td>($0.2148$)</td>
<td>($0.1101$)</td>
<td>($0.1109$)</td>
<td></td>
</tr>
<tr>
<td>$Time^{\text{Post-event Indicator}}$</td>
<td>0.0110</td>
<td>0.0225</td>
<td>0.0235</td>
</tr>
<tr>
<td>($0.0267$)</td>
<td>($0.0137$)</td>
<td>($0.0138$)</td>
<td></td>
</tr>
<tr>
<td>$N1 (User-Event-Days)$</td>
<td>383,040</td>
<td>3,047,240</td>
<td>2,982,980</td>
</tr>
<tr>
<td>$N2 (Users)$</td>
<td>1,368</td>
<td>14,442</td>
<td>14,162</td>
</tr>
</tbody>
</table>

$H_{1a}$ describes effects among our “Top Listener” sample. $H_{1b}$ describes effects among friends of our “Top Listener” sample who are themselves not in the “Top Listener” sample. $H_{1c}$ describes effects among friends of our “Top Listener” sample who have not listened to hate music previously.

**Table 3.10:** Impact of unanticipated events on non-hate music listenership

<table>
<thead>
<tr>
<th></th>
<th>$H_{1a}$</th>
<th>$H_{1b}$</th>
<th>$H_{1c}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Time^{\text{Post-event Indicator}}$</td>
<td>-0.0514</td>
<td>-0.0178</td>
<td>-0.0180</td>
</tr>
<tr>
<td>($0.0122$)</td>
<td>($0.0052$)</td>
<td>($0.0052$)</td>
<td></td>
</tr>
<tr>
<td>$Post-event Indicator$</td>
<td>0.2916</td>
<td>-0.2125</td>
<td>-0.2084</td>
</tr>
<tr>
<td>($0.1571$)</td>
<td>($0.0665$)</td>
<td>($0.0669$)</td>
<td></td>
</tr>
<tr>
<td>$Time^{\text{Post-event Indicator}}$</td>
<td>0.0300</td>
<td>0.0253</td>
<td>0.0250</td>
</tr>
<tr>
<td>($0.0195$)</td>
<td>($0.0083$)</td>
<td>($0.0083$)</td>
<td></td>
</tr>
<tr>
<td>$N1 (User-Event-Days)$</td>
<td>842,688</td>
<td>6,866,160</td>
<td>6,718,824</td>
</tr>
<tr>
<td>$N2 (Users)$</td>
<td>1,368</td>
<td>16,710</td>
<td>16,397</td>
</tr>
</tbody>
</table>

$H_{1a}$ describes effects among our “Top Listener” sample. $H_{1b}$ describes effects among friends of our “Top Listener” sample who are themselves not in the “Top Listener” sample. $H_{1c}$ describes effects among friends of our “Top Listener” sample who have not listened to hate music previously.
Table 3.11: Impact of unanticipated and planned events on non-hate music listenernship

<table>
<thead>
<tr>
<th></th>
<th>$H_{1a}$</th>
<th>$H_{1b}$</th>
<th>$H_{1c}$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Time</strong></td>
<td>-0.0175</td>
<td>-0.0147</td>
<td>-0.0143</td>
</tr>
<tr>
<td></td>
<td>(0.0167)</td>
<td>(0.0045)</td>
<td>(0.0045)</td>
</tr>
<tr>
<td><strong>Post-event Indicator</strong></td>
<td>0.2842</td>
<td>-0.1288</td>
<td>-0.1315</td>
</tr>
<tr>
<td></td>
<td>(0.2148)</td>
<td>(0.0576)</td>
<td>(0.0579)</td>
</tr>
<tr>
<td><strong>Time*Post-event Indicator</strong></td>
<td>0.0110</td>
<td>0.0244</td>
<td>0.0246</td>
</tr>
<tr>
<td></td>
<td>(0.0267)</td>
<td>(0.0072)</td>
<td>(0.0072)</td>
</tr>
<tr>
<td><strong>N1 (User-Event-Days)</strong></td>
<td>383,040</td>
<td>9,913,400</td>
<td>9,701,804</td>
</tr>
<tr>
<td><strong>N2 (Users)</strong></td>
<td>1,368</td>
<td>16,756</td>
<td>16,443</td>
</tr>
</tbody>
</table>

$H_{1a}$ describes effects among our “Top Listener” sample. $H_{1b}$ describes effects among friends of our “Top Listener” sample who are themselves not in the “Top Listener” sample. $H_{1c}$ describes effects among friends of our “Top Listener” sample who have not listened to hate music previously.

3.9.6.2: Event Placebo

3.9.6.2.1: List of Placebo Events

11/21/2015 Louisiana gubernatorial runoff

2/2/2016 Ted Cruz wins Iowa

3/12/2016 Guam, DC, and Wyoming primaries result in Trump losses

2/25/2017 Special Election in Delaware Senate

9/26/2017 Special Elections in Florida Senate and New Hampshire House of Representatives

12/12/2017 US Senate Seat in Alabama Called for Democrat
Figure 3.18: Impact of placebo events on hate music listenership of friends of far-right users

Impact of "Placebo" Events on Hate Music Listens (in tracks played per day)

- Least Active/Central
- Most Active/Central

How Active/Central is the Far-Right Friend?

- Friend Likely Inactive in Far-Right Music Scene
- Friend Likely Active in Far-Right Music Scene

<= Hate Bands than Mean
<= Times a Top Listener than Mean
<= Central than Mean

N=6,479
N=10,038
N=11,523
N=2,738
N=1,209
N=726
3.9.6.3: User Placebo

**Figure 3.19:** Impact of Trump-related events on hate music listenership of non-US friends of far-right users

![Graph showing impact of Trump-related events on hate music listenership](image)

- **N=2,662**
- **N=5,198**
- **N=10,337**
- **N=4,917**
- **N=3,262**
- **N=1,580**

**Legend:**
- ◢ Friend Likely Inactive in Far-Right Music Scene
- ▢ Friend Likely Active in Far-Right Music Scene

**Impact of Trump-Related Events on Hate Music Listens in Placebo Countries (in tracks/day)**

- **Least Active/Central**
- **Most Active/Central**

**How Active/Central is the Far-Right Friend?**

- <= Hate Bands than Mean
- <= Times a Top Listener than Mean
- <= Central than Mean
- > Hate Bands than Mean
- > Times a Top Listener than Mean
- > Central than Mean
3.9.7: **Original PAP**

This PAP was originally posted on [https://osf.io/z5g27/](https://osf.io/z5g27/) but is reproduced here for convenience.

3.9.7.1: **Overview**

This study aims to analyze time-series data to isolate the impact of Trump’s victories and other highly publicized Trump-related events on hate music listenership. While many pundits and political experts have remarked that Trump’s xenophobic rhetoric has revitalized and normalized hate groups in America, to this day, there have been no studies causally identifying that impact in a quantitative empirical study outside of the lab. The research design contained in this document outlines a means of capturing a salient metric of subscription to hate groups and far-right ideologies. We propose a design that will rigorously evaluate the impact of various public events involving Trump (which includes primary and general election victories, noteworthy rallies, and other highly public events) on hate music listenership.

3.9.7.2: **Background**

Pundits and news outlets have remarked that Trump’s presidency has normalized public displays of racism (e.g. see Bass, 2018; Raghunathan, 2018), but there is still a paucity of empirical evidence that causally links Trump’s electoral victories and public statements with increases in racist and discriminatory beliefs, speech, and behavior. One study looked at individual behavior in financially-incentivized games immediately before and immediately after the 2016 election, finding that participants were less likely to cooperate after the election and men appeared to adopt more aggressive strategies towards women (Huang and Low, 2017). Another study by Crandall, Miller & White (2018) looked at a survey of attitudes
from Trump and Clinton supporters before and after the election. They found that the acceptability of prejudice towards targeted groups (e.g. Muslims, Latinos) increased but did not increase for other unpopular groups (e.g. alcoholics, atheists). Additionally, Southern Poverty Law Center reports a spike in hate crimes immediately after the election (Southern Poverty Law Center, 2016). But a critic could claim that racism has always been a problem in America; the Trump presidency may have simply led the media to highlight racially-motivated crimes that were unreported in the past (or to interpret behaviors as hate-related that previously would not have been seen as hate). We propose a means of causally identifying the impact of Trump’s victories and public statements on the activation of racist and xenophobic ideologies by leveraging a novel dataset.

Trump has received an overwhelming wave of support from neo-Nazi and other white supremacist groups like the KKK (Hooton, 2016; Oppenheim, 2017). Further, it has been argued that his reluctance to denounce these groups has led to a resurgence white supremacistism in America (e.g., Smith, 2017, Raghunathan, 2018). To address this question rigorously, we propose examining the actions of these communities that occur immediately after a public, Trump-related event. In many cases, looking at a simple pre/post comparison is insufficient, because there may be other confounding variables that are driving the correlation. One means of causally identifying the impact of a Trump-event on hate group behavior is by leveraging a research design that focuses specifically on immediate effects within a potential outcomes framework. In many cases, this kind of design cannot be implemented because there are very few behaviors that occur immediately after a highly public event. White supremacist rallies must be organized in advance—even hate crimes are unlikely to happen immediately after a white supremacist learns that Trump won a pivotal
primary election. However, here we identify a behavior that does occur on a much more continuous basis: music consumption.

Music consumption is particularly important for hate groups, as it has been used extensively by neo-Nazis and white supremacists to organize and recruit youths (Messner, Jipson, Becker, & Byers, 2007; Corte, & Edwards, 2008, Shekhovtsov, 2013). Indeed, some scholars claim hate music scenes are a primary mechanism for new member recruitment (e.g. see Corte & Edwards, 2008). As such, music by hate bands is a crucial part of many members’ day-to-day behavior. And some of that behavior is available in a publicly available dataset collected via the last.fm API. To clarify, we are not claiming that simply listening to hate music necessarily translates to hate crimes or hate speech. However, subscription to hateful/far-right ideologies is a necessary pre-condition for certain socially harmful behavior (such as “ideologically motivated violence” (Adamczyk, Gruenewald, Chermak, and Freilich, 2014)). We are attempting to estimate subscription to hate ideologies by using hate music listenership as a proxy.

Last.fm is one of the largest public repositories for music consumption information (Henning, & Reichelt, 2008). Although the majority of listened-to-music comes from the most popular artists, there is an insular community of far-right users who listen to hate bands. By looking at their public listenership65 behavior over time, we are able to use public announcements of Trump-related events as a series of exogenous treatments in an interrupted time series design. This would allow us to measure the impact of, for instance,

---

65 Whether users view the tracking of song listens on last.fm as public behavior is not wholly clear. Since last.fm accounts are usually not personally identifiable, there is a level of anonymity to listenership. However, some users do include enough personally identifiable information to match to their social media accounts. Some users may be aware that their listenership is publicly available on last.fm; others may not remember that they enabled the app on their Spotify account.
Trump’s general election victory on hate band song plays. Further, by looking at users who have never listened to hate bands prior to Trump’s electoral successes, we can even test whether the Trump victories either created new recruits or impelled crypto-racists to listen to hate music.

3.9.7.3: Hypotheses

3.9.7.3.1: Primary Hypotheses

H1. We expect that Trump-related public events will increase hate music listenership:

a. among individuals who have listened to hate music extensively in the past (i.e. “Top Listeners” of hate music artists)

b. individuals who have listened to hate music occasionally or not at all (i.e. friends of “Top Listeners” of hate music artists)

c. among individuals who have not listened to hate music in the past (i.e. friends of Top Listeners of hate music artists who have not listened to hate music before any Trump-related event included in this study)

3.9.7.3.2: Secondary Hypotheses

H2. We expect that, for planned Trump-related events, listeners of hate music will increase their listenership of hate music a few days before the event.
3.9.7.3.3: Robustness/Placebo Checks

H3. We expect that both unanticipated and planned Trump-related events will not increase mainstream (i.e. Top 40) music listenership by Top Listeners of hate music.66

H4. Similarly, we expect that Trump-related events will not increase listenership of music similar to hate music by users who are similar to hate music listeners but themselves are not hate music listeners.

3.9.7.4: Dependent Variables

The main dependent variable of interest is daily play counts of a select set of artists by the sample of last.fm users under analysis. This data will be scraped directly from last.fm databases using a custom R wrapper that interfaces with the last.fm API. The list of hate bands is determined as all bands that are tagged as rac, white power, hatecore, rock against communism, white power, nsbm, or ns black metal.67 This yields approximately 6,000 unique bands.68

66 This test helps ensure that far-right users are not simply listening to more music. Rather, they are specifically consuming more far-right music.

67 Last.fm uses a user-based genre tagging system, where every registered last.fm user is able to tag an artist as being in a particular music genre. Many other hate music genre tags have been clearly vandalized and so we refrain from using artists in those tags. Every effort will be made to ensure that there are no cases of artists who are clearly not hate bands included in our hate music list.

68 This number is derived after some manual data cleaning. (For instance, the hate genre tags were associated with hundreds of variants of Russia’s 2008 entry into Eurovision, Dima Bilan—clearly not a hate band.)
Our primary outcome variable will be a daily play count of all hate bands a given user has listened to in the dates ranging from 14 days before to 14 days after every Trump-related event. Cases where an individual appears to have listened to more hours of music than there are hours in a day will be excluded. The secondary dependent variables will be 1) Top 20 artists music listenership, and 2) favorite niche music listenership (i.e. non-hate top artists of friends of hate band listeners).

3.9.7.5: Independent Variables

The independent variable of interest is a set of Trump-related public events. We composed two sets of events 1) unexpected events and 2) planned events. Unexpected events are unanticipated victories, speeches, tweets, or announcements. Planned events are events where the timing is known in advance, so selection effects may be at play. However, any selection effects are likely to bias our estimates downward, since, with planned events, a user could potentially choose to listen to hate music before the event specifically because of that event. While a user could similarly pre-plan to listen to more hate music after the event, it would still be an effect tied to the event. Unless the timing of the event is tied to other points in time that meaningfully impact listenership (e.g. Hitler’s birthday), we can still reasonably claim that the timing of the event is independent of potential outcomes. The list of unexpected events and planned events can be found in the Appendix. We will run three

---

69 Since hate band listeners are likely to listen to other hate bands, for any given user, we scrape not only listens of the band that user is a top listener of, but also all other hate bands in our dataset.

70 Some users may have over a decade of frequent music listenership, so to ensure the scraper is computationally feasible, we must limit our data to as small of a time-period as is reasonable for a time-series analysis. The 14-day window was selected based on findings from prior last.fm research (Ternovski & Yasseri, 2017).

71 With the exception of cases of clear tag vandalism, we will exclude all artists that are found in any two of the following categories: hate music, mainstream music, and favorite niche music.
separate analyses: one that uses only planned events, another that uses only unanticipated events, and a third that uses the combination of the two.

3.9.7.6: Population of Interest

Unfortunately, the structure of last.fm’s API only provides listening data at the level of the individual, rather than total listenership by music artist. As such, even though we can identify hate bands, we are not able to obtain daily aggregate play counts for each artist. We can, however, extract a given artist’s “Top Listeners” or users who have listened to that artist most frequently according to last.fm.\textsuperscript{72} We can thus use our list of hate music to extract top hate music listeners. This procedure yields the sample studied under hypothesis H1a. To construct the sample of individuals examined by H1b, we will also scrape hate band listenership of friends of Top Listeners of hate music. This will ensure that we are sampling not only the fervent far-right, but also more “casual” listeners of hate music. To address H1c, we will use this same sample as in H1b but further limit the sample to only those users who have not listened to any hate music prior to any of the included Trump-related events.

None of these samples are representative and will include users within and outside of the United States. Our analysis will be limited to users who report being in the United States or do not disclose a real country (e.g. users who say they are from Antarctica will be included). A secondary analysis will also examine the effect of Trump events on hate band listenership among users in other countries.

\footnote{72 The top listeners page also requires that the user listened to that artist relatively recently. Unfortunately, the exact timeframe used in the generation of this list has not been published by last.fm.}
3.9.7.7: Analysis

In the current analysis, we are not able to distinguish between people who were and were not exposed to the Trump event. In fact, due to the global reach of social media platforms like Twitter and the extensive media coverage of all Trump-related events, it is likely that relatively few people fail to be exposed to the event in some way at some point. As such, to understand the “control group” behavior of music listenership, we can only do the following: 1) examine whether hate music listeners change their listenership of non-hate music immediately after a Trump event and 2) analyze whether non-hate music listeners who are similar to hate-music listeners change their listenership habits to niche music immediately after Trump events. But these can only serve as placebo control tests, since we are not able to assess how comparable these dependent variables are to the hate-music listenership. As such, we determine that the most appropriate methodology to evaluate whether there is an impact of Trump events on hate music listenership would be a time-series RDD or an interrupted time series. An RDD can be appropriate in this context, as it requires only that the timing is randomly determined and, given continuous outcome data, it would yield a robust estimate of the instantaneous effect (e.g. see Imbens, & Lemieux, 2008). But RDD has rarely been used in time-series contexts, because few effects are truly instantaneous.

A closely related approach is the interrupted time series (ITS) estimator, which is widely used to measure the impact of some event in time on a running outcome (e.g. see Bernal, Cummins, Gasparini, 2017; Briesacher et al., 2013). The average music listenership patterns should be continuous and generally stable over a narrow window of time.

73 The problem in the context of hate music listenership and Trump events is that we do not have a good means of assessing exactly when a user is exposed to the Trump-related event. As such, an instantaneous treatment effect would likely bias the estimate downward.
Exogenous events could have immediate impacts on music choice and if we are able to isolate a class of such events, we can credibly measure an event impact on music listenership. Since hate-music listenership is the consumption of politically radical music, this behavior can be viewed as a low-impact, low-effort political behavior. As such, highly public political events that inspire and/or encourage hate, could mobilize and/or activate low-effort hate-related activities such as the consumption of hate music.

The ITS estimator can be formalized as follows. First, let $Y_{it}$ be the outcome variable of interest for individual $i$ at time $t$. Our primary analysis would look at hate music listenership among two sets of users: HL (top fans of hate bands) and FHL (friends of top fans of hate bands) in the time period of 14 days before and 14 days after a Trump related event. So $Y_{it}$ would be the daily play count of any hate band by individual $i$ at day $t$. Second, we let $D_t$ be a binary treatment variable such that it is equal to one at time $T \geq t_e$, where $t_e$ is the date of Trump-related event $e$. To aggregate the data, the time before and after event $e$ is rescaled as in conventional RDD procedures (Imbens, & Lemieux, 2008). In other words, the cutoff $t_e$ is moved to 0 for every $e$. The timing of an unanticipated Trump event can be thought of as being randomly assigned. This means that if Trump-events do lead to increased hate music listenership, the variable $D_t$ should be positively associated with $Y_{it}$.

Using the standard interrupted-time series framework, we estimate the impact by

---

74 It is arguable that the “unanticipated” definition is not stringent enough; after all, unexpected victories are tied to planned events. For instance, Trump’s Republican primary victory in New Hampshire was inextricably linked to the New Hampshire primary itself—the date of which was made public well in advance of the contest. While Trump’s victory was probabilistic, the date of the contest was not. The crucial question here is whether a hate music listener listens to more hate music because of the victory or because of the contest itself. So long as the choice to listen to more hate music is tied to Trump’s victory and not the event itself, our identification strategy is defensible. As a robustness check, we will determine if Trump losses are also associated with increases in hate music listenership. Additionally, we will check if election contests that do not involve Trump or far-right politicians are associated with increases in hate music listenership.
\[ Y_{t,i,e} = \beta_0 + \beta_1 * T_{t,i,e} + \beta_2 * D_e + \beta_3 * T_{t,i,e} + \beta_4 * D_e + \beta_5 * T_{t,i,e} + \beta_6 * D_e * T_{t,i,e} + \beta_7 * T_{t,i,e} + \beta_8 * D_e * T_{t,i,e} + \epsilon_{t,i,e} \]

where \( T \) is the running time variable, \( \xi \) are the event fixed effects, and \( \zeta \) are the individual fixed effects.

To address the question in H2, we evaluate whether hate music listeners anticipate Trump-related events and listen to hate music before the event.\(^{75}\) We expected planned events would have anticipation effects (i.e. a hate music listener will consume hate music leading up to the event).\(^{76}\) However, even our unanticipated events list may still exhibit anticipation effects (i.e., there are predictions in the media or a user might have the strong prior belief that a particular primary contest will yield a Trump win). To quantify these effects, we will use the following method. We will first randomly select a holdout sample, \( H \). In this sample \( H \), we will shift the rescaled treatment cutoff \( t \) back until the estimate is no longer statistically significant. This would give us a sense as to how many days before the event users begin to modify their listenership behavior in anticipation of the event. We will test this new cutoff in the test data \( T \), where \( T \Theta H \).

To disentangle whether the increase is due to the event itself or some unobserved confound, we pursue a series of placebo control tests. The same methodology as used to evaluate the H1 hypotheses is applied to placebo control data but with the expectation that there are no differences between the observed \( Y_{t,i,e} (1) \) and the inferred \( Y_{t,i,e} (0) \). Namely, we have a similar outcome variable \( U_{t,i} \) (listenership of non-hate music by individual \( i \) at time \( t \)). This variable can be one of the following variants: 1) listenership of mainstream music by

\(^{75}\) Such anticipation effects have been observed in related research on last.fm music consumption (Ternovski & Yasseri, 2017).

\(^{76}\) This would bias any estimates of the event’s impact downward.
individuals who are also top fans of hate bands (hypothesis H3), and 2) listenership of favorite niche music by friends of users who listen to hate music but do not themselves listen to hate music (hypothesis H4). Since the sample in H4 is rather complicated it is useful to formalize: let $i$ be any user that has listened to hate music in our dataset. Let $j$ be any user who is friends with user $i$, but who has not listened to any hate music in any of the time frames of interest. For each user $j$, identify their top listened to artists $A$. We expect that $A$ listenership will not increase after Trump-related events for users $j$. Since last.fm users exhibit high levels of interest-based homophily (e.g. see Bisgin, Agarwal, & Xu, 2012), we expect this population to be most similar to individuals who listen to hate music.

3.9.7.8: Full List of Planned Trump Events

December 7, 2015 – South Carolina rally, calling for “total and complete shutdown of Muslims entering the United States until our country’s representatives can figure out what the hell is going on.” (Trump, 2015 qtd. in Finnegan & Barabak, 2018)


July 19, 2016 – Trump officially awarded Republican nomination (Rafferty, 2016)

July 21, 2016 – Trump speaks at RNC (Plumer, 2016)

August 9, 2016 – North Carolina rally, Trump “suggesting violence against Mrs. Clinton or liberal jurists” (Corasaniti & Haberman, 2016)

77 In this case, niche music is defined as whatever artists that user has listened to most frequently.

78 We use a loose definition of a “friend” connection. Any user that is follower or is following another user is considered in the “friends network” of that user.
September 26, 2016 – Second Presidential debate, Trump says that if he were in charge of the law Clinton “would be in jail” (Roberts, Jacobs, & Siddiqui, 2016)


August 15, 2017 – Press conference, Trump on the rioting from the pro-Confederate rally in Charlottesville: “I think there is blame on both sides….You also had people that were very fine people on both sides….Not all of those people were neo-Nazis, believe me. Not all of those people were white supremacists by any stretch.” (Trump, 2017 qtd. in Finnegan & Barabak, 2018)

August 22, 2017 – Phoenix rally, Trump on the removal of Confederate monuments:

“They’re trying to take away our culture. They’re trying to take away our history. And our weak leaders, they do it overnight. These things have been there for 150 years, for a hundred years. You go back to a university and it’s gone. Weak, weak people.” (Trump, 2017 qtd. in Finnegan & Barabak, 2018)

September 22, 2017 – Alabama rally, Trump on the black football players protesting racial discrimination during National Anthem: “Wouldn’t you love to see one of these NFL owners, when somebody disrespects our flag, to say, ‘Get that son of a bitch off the field right now. Out. He’s fired. He’s fired!’” (Trump, 2017 qtd. in Finnegan & Barabak, 2018)

3.9.7.9: Full List of Unanticipated Trump Events

February 28, 2015 – CNN Interview in an answer about David Duke and the KKK: “Well, just so you understand, I don’t know anything about David Duke, OK? I don’t know anything about what you’re even talking about with white supremacy or white supremacists. So I don’t know.” (Trump, 2015 qtd. in Finnegan & Barabak, 2018)
June 16, 2015 – Tweet: “When Mexico sends its people, they’re not sending their best. They’re not sending you. They’re not sending you. They’re sending people that have lots of problems….They’re bringing drugs. They’re bringing crime. They’re rapists, and some, I assume, are good people.” (Trump, 2015 qtd. in Finnegan & Barabak, 2018)

November 22, 2015 – “Trump retweets fake, racially charged crime data from non-existent group” (Bradner, 2015)

February 9, 2016 – Won New Hampshire Primary (Balz, Eilperin, Fahrenthold, 2016).

February 20, 2016 – Won South Carolina Primary, “clears path to nomination” (Jacobs, Bixby, Siddiqui, Sullivan & Gabbatt, 2016)

February 23, 2016 – Won Nevada Caucus, “further solidifies standing as the front-runner for the GOP nomination” (Goldmacher, 2016)

February 28, 2016 – “Donald Trump Retweets Post with Quote from Mussolini” (Haberman, 2016)

March 1, 2016 – Super Tuesday, “With a big Super Tuesday, Trump has the Republican nomination in his sights” (Barabak, 2016)

March 15, 2016 – Trump wins Florida, Rubio drops out. (Mazzei, Sherman, & Clark, 2016)

May 3, 2016 – Ted Cruz drops out of race leaving “Donald Trump as the only candidate capable of clinching the nomination outright” (Glueck & Goldmacher, 2016)

May 26, 2016 – “Trump reaches delegate count needed to clinch Republican nomination” (McCarthy, 2016)
June 5, 2016 – CBS interview, calling U.S. District Judge Gonzalo Curiel “a member of a club or society very strongly pro-Mexican” (Trump, 2016 qtd. in Finnegan & Barabak, 2018)

July 2, 2016 – Trump retweets anti-Clinton graphic previously posted “on an anti-Semitic, white supremacist message board” (Diamond, 2016).

July 5, 2016 – Trump tweets about “setting the record for the most GOP primary votes ever” (Doran, 2016).

October 7, 2016 – Publication of recording where “Trump brags about groping women” (Graham, 2016).

November 8, 2016 – Trump wins Presidential Election “in stunning upset over Clinton” (Tumulty, Rucker, & Gearan, 2016)


November 29, 2017 – Trump retweet of far-right Nationalist who was “recently arrested for inciting hatred and violence against Muslims.” (Finnegan & Barabak, 2018)
December 23, 2017 – New York Times report claims Trump said Haitians “all have AIDS” and Nigerians “[o]nce they had seen the United States… would never `go back to their huts' in Africa” (Shear & Davis, 2017; Finnegan & Barabak, 2018)

3.9.7.10 Works Cited


3.9.7.11: Works Consulted


3.9.8: Amended PAP

This PAP was originally posted on https://osf.io/z5g27/ but is reproduced here for convenience.

Pre-Analysis Plan Amendment (1/29/2020)

3.9.8.1: Overview

The pre-analysis plan remains at osf.io/z5g27, however due to a new last.fm policy that prevented us from executing the pre-analysis plan exactly as planned, we are registering this amendment. Please note that, as of this date, we have not collected dependent variable data and so we are still blind to the eventual results. This amendment document covers three main modifications: 1) data sources, 2) data cleaning, and 3) two additional hypothesis. Due to last.fm’s new censorship policies, we now make use of data from archive.org. We also take additional steps to confirm that our list of hate music artists is accurate and each band name uniquely identifies a far-right band. Finally, we add two hypotheses to leverage hate music listeners’ friends networks and the amount of hate music they have listened to when analyzing their impact on last.fm friends who have not previously listened to hate music. All three modifications are described in detail below.
3.9.8.2: Data Sources

Last.fm has recently taken action against hate music pages and censored a large subset of artist pages associated with the far-right. In Figure 1, we can see the difference between a censored page and an uncensored artist page. Furthermore, in Figure 2, we can see the effect of this censorship policy on the Listeners pages of both bands. Please note that these changes are also reflected in last.fm’s API.

Figure 1. Uncensored (left) and Censored (right) Artist Pages

Figure 2. Uncensored (left) and Censored (right) Listener Pages
As such, we make use of last.fm’s lengthy history online and archive.org’s Internet archiving project. Using the list of hate bands we had compiled before last.fm’s censorship policy went into effect, we extract all listeners associated with hate bands on archive.org. (Please note, we draw on all relevant last.fm snapshots that are available on archive.org.) Since, currently, there is no distinction between Top Listeners and Recent Listeners, we extract both historic Top Listeners and Recent Listeners. Our original Pre-Analysis Plan explicitly focused on Top Listeners and not Recent Listeners so as to avoid conditioning our inclusion criteria on a dependent variable generated post-treatment (i.e. Recent Listeners have listened to hate music after our analysis window). However, since we can now determine that a last.fm user was a listener to a given hate music band before our analysis window, this concern is no longer relevant, so we can combine Recent and Top Listeners.79 To be explicit, because we now have historic snapshots of hate music listeners, we only include last.fm users in our list of hate music listeners if they listened to hate music before 2015.

Since last.fm did not censor song listens or last.fm friends networks, the rest of our pre-analysis plan need not be modified.

3.9.8.3: Data Cleaning

A closer look at our list of hate bands revealed that despite the precautions we took to avoid vandalized genre tags, there were artists who were clearly mistagged (e.g., j-pop artist, Kahimi Karie) or spurious (e.g., Geico [the Insurance Company]). Additionally, we noted that some hate bands had the same name as artists who had no connection to far right ideologies (e.g., Evil). As such, for every artist where we were able to successfully extract

---

79 Additionally, last.fm never publicized the exact algorithm for calculating Top Listeners, so we no longer need to rely on a proprietary black-box procedure.
listeners from archive.org, we manually looked up the last.fm artist page in a web browser and inspected the genre tags, songs, album titles, the shout-box, the artist biography, pictures of the artists, and related artists. We then hand-coded them as either “mistagged/fake”, “suspect”, and “multi.” Mistagged/fake was used in cases where it was clear that the artist was either not a hate music artist (e.g., several Eurovision pop stars) or not a real music artist (e.g., politicians, cartoon characters, etc.). “Suspect” was used in cases where the connection to far-right ideologies was either disputed (based on the shout-box activity and/or artist biography), the band was a troll artist (e.g., certain grindcore bands), or the ideologies were not immediately clear (e.g., anti-racist hatecore). Finally, an artist was tagged “multi” if the biography or the shout-box made it clear that multiple artists shared the same band name. One important note: there were a few cases where multiple artists shared the same name, but all were explicitly far-right; these cases were not tagged as “multi,” since, for our purposes, they are no different from uniquely named far-right bands.

To reduce the amount of noise in our data, we exclude all artists (and the corresponding listeners) who were tagged as either “mistagged/fake,” “suspect,” or “multi.”

3.9.8.4: Additional Exploratory Hypotheses

We also noted that the collected data allows us to have greater precision when assessing whether friends of far-right music listeners experience increases in hate music listenership following Trump-related events (Hypothesis H1c). Particularly, we can 1) map the network of far-right users and 2) evaluate how much far-right music a given user has listened to in the past. Both data may estimate how radicalized a user is. Specifically, we believe that if we map the network of far-right music listeners, individuals that are more central in the network are likely to be more strongly enmeshed in neo-Nazi and other far-
right music scenes. Similarly, individuals who listen to more hate music (both in terms of time $t$ [i.e., the number of archive.org snapshots] and quantity $q$ of hate bands listened to) should, on average, have stronger affinities to far right ideologies. As such, both types of listeners are likely to be more influential in radicalizing their last.fm friends. Therefore, we add the following two hypotheses:

H1c1: Music listeners who have not listened to hate music in the past should have larger treatment effects if they are friends with active hate music listeners who have higher levels of eigenvector centrality.

H1c2: Music listeners who have not listened to hate music in the past should have larger treatment effects if they are friends with active hate music listeners who have listened to more hate music (in terms of $t$ (number of archive.org snapshots) and $q$ (number of hate bands listened to)).

3.9.8.5: Additional References


195


4.1: Abstract

While there is compelling evidence that the act of casting a vote in one election makes it more likely that the voter will vote again in future elections (Coppock and Green 2016), there is still a lack of clarity as to how and why such a “habit” develops. In this paper, I apply a new theoretical framework, the Process Model for Behavior Change (PMBC) (Duckworth and Gross 2020), to voting to more cohesively bridge the economic cost-benefit model of voting with the psychology-motivated voting-as-a-habit literature. This new theoretical frame gives greater clarity as to how a vote in one election might beget a vote in another election, while yielding testable predictions as to which circumstances are more favorable for developing turnout persistence. This paper also makes use of nine large-N, door-to-door voter mobilization field experiments in various election contexts (~1.8 million voters in total) to evaluate predictions from the PMBC model and from prior empirical research. Consistent with prior empirical research, my analysis finds that being persuaded to vote in one election does lead to increased turnout four years later. However, the PMBC predictions and my empirical research both diverge from the conclusions of several prior empirical studies. For instance, being induced to vote in low salience elections does not necessarily translate to voting in downstream elections. These results illustrate the usefulness of the PMBC theoretical framework to reconcile and make sense of, at times, contradictory empirical results in the voting-as-a-habit literature.
4.2: Introduction

Political science has made tremendous headway in uncovering the mechanism behind why people vote since the canonical formalizations of Riker & Ordeshook (1968) and Downs (1957), who conceptualized the decision to cast a ballot as hinging on a cost-benefit equation. There is now copious evidence that social image also plays a key part in voters’ decision to vote or abstain (e.g., Green, McGrath and Aronow, 2013; Rogers, Ternovski & Yoeli; 2016, Fujiwara, Meng, and Vogl, 2016; DellaVigna, 2016; Gerber et al., 2016). This literature has a well-developed theoretical framework and strong empirical evidence validating many aspects of the theory. There has also been a branch of empirical studies investigating if voting is “habit-forming.” (Solvack and Vassil, 2018; Coppock & Green, 2016; Garcia Bedolla & Michelson, 2012; Franklin & Hobolt, 2011; Meredith, 2009; Gerber, Green & Shachar, 2003). Coppock and Green’s (2016) overview of the existing literature and empirical contributions provide compelling evidence that voting in one election does cause an increase in the likelihood that the same individual will vote in a future election. But there are key components of the theoretical framework that are, as yet, undeveloped. For one, a voting “habit” does not satisfy psychologists’ definition of habit as an action that becomes automatic from continuous repetition (Lally and Gardner, 2013)—the long gaps of time between elections preclude voting from becoming truly automatic (Dinas, 2012). But more pressing is the concern that if voter is habit-forming, why are there so many cases where voting in one election does not lead to another vote in a subsequent election? Why do so many voters lapse, while others vote only sporadically? As Coppock and Green (2016) concluded in their empirical contribution and review of existing literature, turnout persistence “is more complex than suggested by prior work in this area.” (p. 1060).
And I argue that complexity demands a cohesive, comprehensive theoretical framework that yields concrete predictions that can be empirically validated.

The paucity of a compelling theoretical framework for turnout persistence is partly explained by the difficulty of reconciling habit with existing models of voter behavior. An economic perspective on habit is too narrow: voting persistence in this context is reduced to a lagged turnout variable in an equation modelling the likelihood of future turnout. This approach fails to provide concrete predictions as to why voting could be habit-forming in the first place.80 On the other hand, the psychological perspective must grapple with the fact that turnout persistence fails to meet the strict criteria psychologists have for habitual behavior. How can voting be habitual, when the behavior not automatic, is not repeated frequently, and there may be years between each election? Fortunately, recent developments in judgement and decision-making have led to a new theoretical framework that can clearly depict the mechanism of seemingly habitual behavior without the need to determine whether a behavior is, strictly-speaking, a psychological habit.

One of the key contributions of this paper is leveraging this new theoretical framework, Duckworth & Gross’s (2020) Process Model of Behavior Change (PMBC), to give greater clarity as to how “habitual” voting relates to economic (i.e., cost-benefit) models of voting. The PMBC model incorporates the classic cost-benefit analysis into a stage of cognitive decision-making (the Appraisal Stage), which can be skipped under the right circumstances through a cognitive shortcut (such shortcuts are usually habits). This model predicts several key factors that should lead to stronger “habitual” voting effects: 1)

---

80 Note that economic models of voter habit explicitly attempt to disentangle habit formation from the “stability over time in the benefits and costs of voting” (Fujiwara, Meng, & Vogl, 2016, p. 161).
contextual cues that bring attention to an election should be similar to that of past elections, 2) past voting experiences should be subjectively positive, and 3) repetition of the first two factors across elections should yield voting patterns that resemble habitual behavior.

The second key contribution of this work is empirical. Through a partnership with a large-scale political organization, I make use of an ~1.8 million voter dataset to evaluate downstream turnout effects that are caused by upstream voting behavior. To the best of my knowledge, this is the largest such experimental evaluation. This data has three key advantages:

1. The sample is predominantly older and allows us to look at the effect of GOTV outreach among a population that has had many previous opportunities to evaluate their individual costs and benefits of voting but failed to become frequent voters. This allows us to better guard against the possibility that it isn’t the act of voting that leads to future turnout, but, rather, it is the large informational bolus from voting for the first time that permanently updates a new voter’s cost-benefit equation of voting.

2. The data includes 8 different states with large sample sizes across 5 election contexts, which allows us to establish a greater degree of generalizability. Additionally, the size of my sample and the diversity of the election contexts also allows me to evaluate which specific variables are conducive to developing habit-like voting behaviors.

3. The field experiments in this paper specifically use GOTV outreach that is designed to foster habitual voting by leveraging plan-making intentions and using consistent
contextual cues across elections. This also helps guard against the possibility that the cost-benefit equation of voting has changed.

In sum, this paper presents a novel theoretical framework from the cognitive sciences and applies it to the puzzle of voting persistence. By applying the PMBC model to previous studies, I am able to generate concrete predictions as to the circumstances of when turnout persistence should be stronger or weaker. I then test these predictions in a large-N dataset. As such, this paper has implications both for 1) public policy by providing some guidance as to how to develop “habitual” voting among an older population and 2) political science as the PMBC model and my empirical results better define the mechanism behind turnout persistence.

This paper is organized as follows. First, I provide a brief summary of the commonly-used cost-benefit model of voting. I then explicate the key issues this type of model has in incorporating turnout persistence (or “habitual” voting) by discussing economic models of habit and how habit is defined by psychologists. I propose that the best way to cohesively bridge these literatures is through Duckworth and Gross’s (2020) PMBC theoretical framework. In the next section, I apply the PMBC model to voting. I then discuss the results from the voting-as-habit literature through the lens of the PMBC model. In the following section, I turn to my empirical work; I describe the data used in this paper and explicate concrete predictions borne out of the PMBC model that I test in this paper. In the final two sections, I present the results, discuss implications, and conclude.

4.3: The Calculus of Voting

To construct a theoretical model of why casting a vote makes an individual more likely to vote in subsequent election, we must first establish the mechanism for why any
citizen votes in the first place. Our understanding of why voters turn out to the polls has
developed considerably since the canonical formalizations of Riker and Ordeshook (1968)
and Downs (1957). Their pioneering work noted that there is some cost of voting and, in
large elections, a single vote is highly unlikely to matter, so what is the benefit of turning out
to vote? Why do so many people still incur the personal cost (e.g., time costs, transportation
costs, etc.) of voting even though it is highly improbable that their vote will decide the
outcome of an election? Downs (1957) wrote of the perceived likelihood of being the pivotal
voter and Riker and Ordeshook (1968) added the “warm glow” variable to the model. In
other words, a voter considers how close the election will be in their mental calculus, but
also anticipates a warm glow after fulfilling one’s civic duty at the polls. And so, the voter
may incur the cost of voting even in elections that aren’t close. Their reliance on the concept
of the “rational voter” and insistence on including a “pivotality” variable (i.e., the expected
probability that one’s vote will decide an election) set off a heated debate between the
Downsian political economists and psychology-influenced political scientists (for summaries
see Blais, 2000; Green and Shapiro, 1996). As such, there was a proliferation of theoretical
models and empirical research that attempted to better understand what the Downsian
school called the “paradox of voting” (for a summary of the relevant approaches and a meta-
analysis of empirical research, see Smets & Van Hem, 2013). But though some elements of
the original model continue to be controversial (i.e., do people really think about the
probability that they can swing a presidential election?), even some of the leading critics of
the rational voter model came to use cost-benefit equations to model an individual’s decision
to vote or abstain in their own work (e.g., see Gerber, Green and Larimer, 2008). Therefore,
as a first step, I assume that a cost-benefit equation is one key determinant of a would-be
voter’s decision-making process.
There is also some consensus as to several of the variables included in that equation: there likely exists some cost of voting and that there may be some *intrinsic* benefit of voting (e.g., Gerber, Green & Larimer, 2008; Meredith, 2009; DellaVigna et al., 2016; Fujiwara, Meng and Vogl, 2016) which we conceptualize as the original “warm glow” variable from Riker and Ordeshook (1968). Critics of the “pivotality” variable (i.e., the variable that denotes the expected probability of casting the vote that swings an election) maintain that it is likely so small that it could be treated as zero (e.g., Gerber, Green, & Larimer, 2008), but because this variable is still included in many economic models of voting (e.g., DellaVigna et al., 2016; Fujiwara, Meng and Vogl, 2016), I include it in my cost-benefit equation with the acknowledgement that it may be zero (for a complete formalization, see DellaVigna et al., 2016).

Aside from considering these well-established cost and benefits of voting, it has become increasingly clear that there is a powerful social element to voting. Gerber, Green and Larimer (2008) experimentally induced prospective voters to turn out via social pressure by publicizing their neighbors’ voting records. The social pressure intervention has been replicated in many other field experiments (for overviews and meta-analyses, see Green & Gerber, 2019; Green, McGrath and Aronow, 2013) and the concept of “social image” was finally incorporated into formal models of voting (e.g., DellaVigna et al., 2016). In other words, there is a widespread belief that voting is an important part of what it means to be a

---

81 This variable is also referred to as the “civic duty” benefit of voting (e.g., Gerber et al., 2016), but since this benefit may imply some extrinsic social dimension (i.e., “duty” implies some obligation to others, which may be enforced through social sanctions), I opt to use “warm glow.”

82 This intervention exploited the fact that whether or not someone has voted is public record in the United States.
good citizen,\textsuperscript{83} so there may be negative social repercussions for abstaining and potentially positive social rewards for turning out to the polls. And indeed, scholars have found that voters relinquish monetary rewards to tell others that they vote (DellaVigna et al., 2016) and pay to find out if others voted (Gerber et al., 2016). This is all to say that there are strong norms that citizens should vote, so people may turn out to vote because they perceive or anticipate some form of real or imagined social sanctions for being a non-voter (Fujiwara, Meng, and Vogl, 2016; DellaVigna, 2016; Gerber et al., 2016).

For the purposes of this paper, I use the DellaVigna et al. (2016) model as the core cost-benefit model\textsuperscript{84} for three reasons: the costs and benefits in this model are commonly used in other formalizations, it has extensive empirical validation backing its assertions, and it is parsimonious. However, this paper does not hinge on the choice of cost-benefit models and alternative models could easily supplant the DellaVigna et al. (2016) model in my analyses.\textsuperscript{85}

In the DellaVigna et al. (2016) model, there are three benefits and two costs. The benefits are

1. the expected utility of being the pivotal voter (i.e., your vote decides the election),

2. the “warm glow” of voting,

\textsuperscript{83} “90% of respondents on the 1984 GSS claimed voting was a very important civic obligation” (Rolfe, 2012, p. 50)

\textsuperscript{84} Though for conceptual clarity in later discussions, I decomposed their “social image” functional into social cost and social benefit.

\textsuperscript{85} In fact, the exact cost-benefit equation can vary from person to person. In the theoretical framework I propose in this manuscript, the cost-benefit equation is simply one stage in an overarching process of recurring behavior. The cost-benefit equation for any given individual outputs an expected utility; the other stages in my theoretical framework use only the resulting expected utility and do not refer to the specific costs and benefits that produced that expected utility.
3. and the social utility of voting (e.g., being able to truthfully report having had voted to peers and family)

whereas the costs are:

1. the transactional cost of voting (time, transportation costs, opportunity cost, etc.),
2. and the social cost of not voting (e.g., either telling the truth and being socially sanctioned for not voting or the cognitive cost of lying about one’s abstention).  

4.4: Reconciling “Habit” with the Calculus of Voting

There may be yet another factor that determines whether any given individual casts a ballot: voting appears to be “habit-forming” (Solvack and Vassil, 2018; Coppock & Green, 2016; Garcia Bedolla & Michelson, 2012; Franklin & Hobolt, 2011; Meredith, 2009; Gerber, Green & Shachar, 2003). It has long been noted that prior voting is an excellent predictor of future voting (e.g., Brody and Sniderman, 1977), but it was not clear if voters were simply running the calculus of voting in their head each time and coming to the same result (i.e., the benefit of voting exceeds the cost of voting) or if voting truly became habitual. Coppock and Green (2016) conclude that the body of empirical evidence “leaves little doubt that voting is habit-forming” (1046), but there are still lingering questions as to the mechanism of habit-formation. For instance, as Dinas (2012) points out, a voting “habit” isn’t really a habit in the psychological-sense. Psychologists commonly define habit as a set of “behavioural patterns enacted automatically in response to a situation in which the behaviour has been performed repeatedly and consistently in the past.” (Lally and Gardner, 2013; 137). Though it is important to emphasize that habits also require “routine contextual cues” in the

---

86 Note that the theory of expressive voting (e.g., Rogers, Fox, and Gerber, 2013) can be incorporated in the “warm glow” variable and the social cost/benefits of voting (e.g., Etang, Fielding, & Knowles, 2016).
environment (Beshears et al., 2020). In other words, an individual receives some cue (e.g., it’s morning and I’m in my car about to go to work) and takes some automatic action (e.g., I take the same route I always take without deliberating or actively thinking about as to which streets to take). In short, a conventional psychological model of habit-formation is not fully compatible with the notion of habitual voting for two reasons. First, habitual actions are “automatic,” which implies an “absence of awareness, conscious control, mental effort and deliberation” (Lally and Gardner, 2013; 137). It is unlikely that even the most consistent of voters are “unaware” going to the polls as one would drive down a familiar route.87 The second issue is that to translate a deliberative behavior to an automatic behavior, substantial repetition is required to achieve any level of automaticity—for instance, a study looking at the adoption of healthy daily behaviors found that the behavior became automatic only after 66 days on average (Lally et al., 2010). Even if someone votes in three elections every two years (e.g., a municipal election, a party primary, and a general election), that voter would only be expected to develop a voting “habit” after 44 years of being a voter.88 This stands in stark contrast to the fact that much of the voting-as-a-habit literature examined much shorter time horizons but still found turnout persistence.

On the other hand, economics has a rich literature of incorporating consumer habit in formal models of economic behavior (e.g., Koford and Miller, 1991; Carroll, Overland & Weil, 2000; Meer, 2002). And from this perspective, Fujiwara, Meng & Vogl (2016) defined voting habit as “voting today, holding constant voters' characteristics, affect[ing] voting decisions in the future” (Fujiwara, Meng & Vogl, 2016; p. 165). To avoid confusing habit in

87 For a more thorough criticism of construing persistent voting as “habitual,” see Dinas (2012; 435-436)

88 (2 years / 3 votes) X 66 days
a psychological sense with this narrow economic definition of habit, I adopt the term proposed by Green & Shachar (2000): consuetude, defined as “merely engaging in the activity makes it more likely that one will engage in the same activity.” (p. 562). As such, in Fujiwara, Meng & Vogl (2016), voting consuetude is essentially formalized as a lagged outcome variable (i.e., whether or not a given individual voted in the last election). But the issue with this economic definition is that it is too narrow to yield concrete predictions as to why the empirical studies found voting consuetude in some contexts but not in others (for a brief overview of some of these inconsistencies, see Coppock and Green, 2016).

In short, these two strands of literature are both incomplete in their treatment of voting consuetude. On one hand, the economic model is precise but so narrow that it, in itself, provides no guidance as to the mechanism behind why one vote is more likely to lead to more votes. The psychological model of habit, on the other hand, has concrete predictions as to how habits form (for a brief overview, see Lally & Gardner, 2013), but is clearly inapplicable to an activity that may occur as rarely as once every few years and is clearly never “automatic” in the psychological sense.

Fortunately, Duckworth and Gross (2020) present a more general decision-making model that can better be reconciled with persistent voting behaviors. Specifically, Duckworth and Gross’s (2020) Process Model of Behavior Change (PMBC) is a theoretical

---

89 “Consuetude is conventionally defined as habit or custom but lacks the unwanted connotations of those terms. The term ‘habit’ calls to mind such activities as cigarette smoking or drug addiction, in which a person is locked into a pattern of conduct by forces that are in some sense outside his or her control. Similarly, to call voting a ‘customary activity’ directs more attention than we would like to the effects of the cultural context in which voting occurs. Absent from common parlance, consuetude provides an empty vessel into which we may pour… meaning.” (Green & Shachar, 2000; p. 561-562)

90 Studies that use such formal models of voting tend to draw on other theories and empirical research for concrete predictions (e.g., Fujiwara, Meng & Vogl, 2016).
framework\footnote{Duckworth and Gross (2020) themselves note that the “Process Model of Behavior Change” is actually a theoretical framework rather than a model in the strict sense, but since they use the PMBC shorthand, I use “model” and “theoretical framework” in this paper interchangeably.} for analyzing behavior change that incorporates both non-habitual activities and habitual activities. But crucially, it does not require for an activity to be a habit in the strict-psychological sense. Instead, habit is just one of many types of heuristic decision-making processes that are incorporated in the model.\footnote{An example of heuristic decision-making that is not habitual would be entering a new cafeteria and opting for a new healthy food option, because it is most prominently displayed (e.g., Just & Gabrielyan, 2018).} That said, it is important to emphasize (as Duckworth and Gross (2020) themselves do) that these features are not novel and they have all been discussed and studied extensively by many scholars in the past. Rather, this model is an integration and distillation of existing behavioral research; its main goal is to provide a parsimonious yet rich framework with which to analyze behavior change (Duckworth and Gross, 2020; p. 45). Much like the Downsian scholars that attempt to reduce complex cognitive process behind the decision to vote into simple testable models, the Duckworth and Gross (2020) theoretical framework provides a similarly succinct encapsulation of key cognitive processes behind individual behavior during repeated activities.\footnote{My application of this model to voting consuetude follows in that tradition. All my analyses of mechanism through the lens of this model have been touched on in the past by other scholars in a more ad hoc way (e.g., Rogers & Frey, 2014); my motivation for using the PMBC is to generate concrete testable predictions from a coherent, self-contained theory and to evaluate those predictions in a large empirical dataset.}

In sum, the reason why the PMBC is particularly useful to gaining a better understanding of voting consuetude are as follows. First, unlike narrow definitions of voting consuetude, the PMBC provides clear, testable predictions as to which circumstances are more favorable for turnout persistence. Second, unlike the theories of habit formation that are cited extensively by the voting-as-habit literature (e.g., Solvack and Vassil, 2018; Aldrich, Montgomery & Wood, 2011) the PMBC avoids the paradox of how such an infrequent and
clearly non-automatic behavior can be construed as habitual in the strict psychological sense.

Third, the PMBC is a parsimonious, rich, and coherently self-contained theoretical framework. This avoids cherry-picking among many competing and often over-lapping theories (for a critique of this commonly-used approach, see Smets and van Ham, 2013). As such, applying the PMBC to voting consuetude allows us to reconcile inconsistencies in the voting-as-a-habit literature. Namely, it is not uncommon for empirical studies to find contradictory evidence. For instance, some studies found that previous voting in one type of election does not translate to downstream voting in a different type of election (Michelson, 2003; Hill & Kousser, 2016), but other studies found voting consuetude effects even for different types of elections (Garcia Bedolla and Michelson, 2012; Gerber, Green, & Shachar 2003). In the next section, I discuss the Duckworth and Gross (2020) model and how it applies to voting consuetude to illustrate how this model can help make sense of the sometimes-contradictory results from the voting-as-a-habit literature.

4.5: Skipping the Calculus of Voting; Voting as a “Habit”

The PMBC stipulates that any behavior occurs as part of a recursive cycle of four stages. A graphical depiction of this model as applied to voting can be seen in Figure 4.1 (left). First, we encounter a situation—for instance, it is Election Day. Some element or elements of the situation demand our attention (such as seeing news coverage of the upcoming election). We appraise the situation (i.e., essentially apply our own personal cost-benefit equation to determine if voting is worth the effort) and pick the appropriate response (i.e., vote or not). But this model also emphasizes that because appraisal is cognitively taxing, there exist shortcuts past the appraisal stage (see Figure 4.1 right). To be specific, this cognitive shortcut economizes “on cognitive effort and are enacted not because we calculate that their net benefits minus costs are optimal in the moment but rather because we have
responded the same way in the same context and gotten a similar reward” (Duckworth and Gross 2020, 41). Past voting causing future voting would therefore be described as a sidestepping Appraisal Stage and, therefore, not weighing the costs and benefits of voting.

**Figure 4.1**: PMBC model as applied to voting before a voting “habit” (left) and after a voting “habit” develops (right)

![PMBC model diagram](https://via.placeholder.com/150)

*Adapted from Duckworth and Gross (2020, p. 40)*

While repetition is undoubtedly necessary to make voting truly habitual, because the PMBC is a process-driven model, we can focus on the mechanism behind the development of shortcuts past the Appraisal Stage and we do not need to adjudicate whether persistent voting is really a “habit” in the strict psychological sense. Particularly, the cognitive shortcut does not have to imply that the Response is automatic in the sense that the initiation of the action is not consciously recognized. Rather, one can construe this shortcut as simply a failure to deliberate over whether a given action is worth taking. In other words, the cognitive shortcut could simply be a line of reasoning such as “Last time I voted, I was glad that I voted, so I may as well vote again.” This is distinct from a deliberative thought process that has the individual thinking about how much time they’ll wait in line at the polling place,
the times someone praised them for voting, etc. In other words, the cognitive shortcut could simply be a failure to even cursorily engage in the specifics of a cost-benefit assessment. And so, the PMBC model allows us to identify what elements are necessary for voting to persist in a manner resembling a habitual action (i.e., voting without appraisal as to whether one should vote or not). I delineate these necessary conditions below:

1. An individual’s attention must be triggered by some environmental cue about election day (i.e. the Attention Stage).
2. Previous voting experience must validate that the voting calculus in the appraisal stage was correct. (In other words, the cost-benefit calculation should clearly show that the voter made the “right” choice in showing up at the polls.)
3. There must be continuous positive reinforcement as the voting behavior is repeated\(^{94}\) for the voter to begin to skip the Appraisal Stage and show up at the polls as a cognitive shortcut.

We can now use the theoretical framework to re-evaluate existing empirical research on voting consuetude, which I do in the next section.

4.6: “Voting-as-a-Habit” Literature and the PMBC Model

There have been two major approaches to isolating voting consuetude from confounding variables: 1) a regression discontinuity at the point when a voter becomes eligible to vote, 2) a 2SLS regression on downstream turnout among compliers in

---

\(^{94}\) As Dinas (2012) points out, voting in the upstream election and voting again the same year in a downstream election should be not construed as a habit, so a longer time horizon is necessary to evaluate whether a shortcut has formed.
randomized (e.g., GOTV nudges) or natural experiments (e.g., rainfall). I discuss each of these approaches in turn and apply the PMBC model to make sense of their findings.

The first approach is a temporal regression discontinuity design that uses the date at which a voter becomes eligible to vote (Meredith, 2009; Franklin & Hobolt, 2011; Dinas, 2012; Coppock and Green, 2016). In other words, this approach compares the long-term vote record of the voter who turned 18 just before an election to the long-term vote record of the voter who turned 18 just after the same election. This design exploits the fact that people cannot self-select to either group and assumes that the only meaningful difference between the two groups is eligibility to vote. With this approach, researchers have found that voting habits persist for at least twenty years (Coppock and Green, 2016), which is consistent with the broader psychological literature on habit-formation—repeated habitual actions should be self-reinforcing (e.g., Lally and Gardner, 2013).

To better see how this research would be interpreted by the proposed PMBC model, it is useful to explicate each step for a young voter who becomes eligible to vote shortly before an election. Under the PMBC model the Situation Stage is the upcoming election that a young voter will be eligible for. Next, that young voter may be made aware of the upcoming elections via campaign and non-profit outreach, media coverage, and conversations with friends and family (i.e., the Attention Stage). If they are not made aware of the upcoming election, they will therefore abstain without ever assessing the costs and benefits of voting. If they are made aware of the upcoming election that they are eligible to vote in, the young voter will proceed to the Appraisal Stage, since they have not voted before and, thus, have no cognitive shortcuts past the Appraisal Stage. As mentioned earlier, for the Appraisal Stage, I use a variant of DellaVigna et al.’s (2016) model, which has three
benefits and two costs. The benefits are the expected utility of being the pivotal voter (i.e., your vote decides the election), the “warm glow” of voting, and the social utility of voting (e.g., being able to truthfully report having had voted to peers and family). The costs are the transactional cost of voting (time, transportation costs, opportunity cost, etc.) and the social cost of not voting (e.g., either telling the truth and being socially sanctioned for not voting or the cognitive cost of lying about one’s abstention).

What is important to note is that the new voters must estimate these variables with limited information and those estimates may be wildly inaccurate. Additionally, these variables can be manipulated exogenously. For instance, social pressure GOTV outreach explicitly (e.g., Gerber, Green and Larimer, 2008) or implicitly (e.g., Rogers, Ternovski and Yoeli, 2018) emphasizes that whether or not someone voted is public record and this tactic has consistently increased turnout even in competitive elections (e.g., Green, McGrath and Aronow, 2013). This is to say that, under this theoretical model, receiving a piece of mail that emphasizes that whether or not you voted is public record and people may ask you about it is likely to increase the expected cost of not voting. The key takeaway is that social and campaign outreach that calls attention to an upcoming election can also affect the perceived variables in the cost-benefit model used in the Appraisal stage of the PMBC decision-making model. Hence, the young voter in our example would either choose to vote or abstain and proceed to the Situation Stage of the PMBC model. In this example, the Situation Stage can be construed as encompassing both the experience at the polls and the subsequent social interactions about voting.95

95 For a more detailed discussion of how social interactions after voting can exacerbate voting consuetude, see Fujiwara, Meng, and Vogl (2016; 182-3).
At the Situation Stage, the voter will therefore get important information that updates the values of each variable used in the cost-benefit analysis, but the voter and the non-voter have systematically different sets of information. For one, the voter can update the “warm glow” variable depending on one’s experiences at the polls. Someone who had a positive experience may assign a higher value to their “warm glow” benefit, while someone who waited in a long line and was affected by voter ID laws (e.g., they cast a “provisional” ballot because their identification did not meet state requirements), may decrease the value assigned to the “warm glow” variable.96 The non-voter would thus only be able to update this variable through second-hand information (e.g., talking to friends and family who voted). And as Meredith (2009) points out, the voter may also be able to reduce the transaction cost of voting by getting important first-hand logistical information about voting (e.g., knowing where to go, what the ballot looks like, how voting machines work, knowing more about the candidates). The non-voter would not get the benefit of this cost reduction. The voter and non-voter will get markedly different pieces of information that affect the social cost and social benefit of voting. The non-voter will get more information about the social cost of not voting and the voter will get more information about the social benefit.

On the next election day (which may be years later), our example youth voter encounters a similar situation and must again get some reminder to bring their attention to the election. If they are made aware of the election they may again proceed to the Appraisal stage with updated cost-benefit values, which may be updated again in a similar way. Ultimately, this updating of cost-benefit terms can also be considered in a reduced form: was

96 It’s worth noting that there are some scholars who asserted that the warm glow variable from voting is so large that it doesn’t lead to voting consuetude, but is also “transformative” in changing individuals’ overall political engagement (e.g., Jakee and Sun, 2006). A large meta-analysis has found little evidence that voting is transformative (Holbein et al., 2021).
the decision to vote or not vote “correct”? In other words, in casting a vote did the voter feel that the benefits justified the costs and vice versa. This aspect is critical in habit-formation because a desirable behavior (i.e., voting) only starts to become a habit if there is a positive benefit to that behavior and the contextual cues drawing attention to the upcoming election are similar (Duckworth and Gross, 2020). At some point, an individual can therefore sidestep the Appraisal Stage because the cost-benefit maximization equation has led to positive results in the past.

The results from this RDD literature could imply that over the course of multiple elections, newly eligible voters develop a persistent heuristic-driven shortcut past the Appraisal stage. But there does remain one alternative explanation that I alluded to earlier: divergences in the voting record of the voter who became eligible to vote just before an election versus that of the voter who didn’t may be driven by divergences in the informational boluses the two types of individuals receive (i.e., this is an exclusion violation). The key part is that newly eligible voters may have highly inaccurate estimates of the costs and benefits of turning out to the polls and voting for the first time could potentially provide a major correction to a misestimated cost or benefit. The amount of information a new voter receives from voting in each successive election is likely diminishing on average. As such, to better disentangle whether it is the act of voting or the information received from the voting experience that is responsible for future votes, it would be useful to examine the development of heuristic shortcuts among more seasoned voters.

Some of the empirical research in the voting-as-a-habit literature does indeed make use of randomized controlled experiments to include the development of voting consuetude among older voters (e.g., Coppock & Green, 2016). This population is composed of people
who have already made it a habit to abstain in elections or those individuals who have gotten a good sense of the costs and benefits of voting but did not become frequent voters. It is worth emphasizing that the PMBC model was constructed to gain a better understanding of how to modify existing, undesirable behaviors and supplant them with healthy, desirable behaviors (e.g., healthier eating choices) (Duckworth and Gross, 2020), which makes the PMBC model particularly suitable for studying the long-term implications of inducements to vote among a broader population of individuals who have “unhealthy voting habits.”

For this population of infrequent voters, the voting-as-a-habit literature makes use of 2SLS regression to look at the long-term vote histories of individuals who were induced to vote due to random or as-if random interventions. Most of this literature using this approach made use of randomized field experiments that were originally conducted not to build a voting habit but to simply get voters to vote in a particular upstream election (e.g., Gerber, Green, Shachar, 2003; Michelson, 2003; Garcia Bedolla and Michelson, 2012; Coppock and Green, 2016). Among the most successful of these interventions is the social pressure nudge (Green, McGrath, and Aronow, 2013) and this outreach explicitly targets the social cost and social benefit terms in the cost-benefit equation at the Appraisal Stage.

While these experiments analyze a more experienced voting base and therefore likely avoid the problem of large informational boluses from turning up at the polls for the first time, they may have a different exclusion restriction violation. As Fujiwara, Meng, and Vogl (2016) note, it is possible that the observed downstream effects of these nudges are due to lasting changes in the cost-benefit equation during the Appraisal Stage as opposed to a cognitive shortcut past the Appraisal Stage. To be clear, social pressure nudges aim to increase both the social cost and social benefit of voting by emphasizing that whether or not
someone voted is public record. The treated individual may thus anticipate more people asking about whether they voted, but after a few elections they may realize that being asked about voting is not as frequent a situation as they expect, which may lead to cessation of voting after just a few elections. And indeed, this is what Coppock and Green (2016) see in their analysis of downstream voting among compliers of GOTV field experiments, which stands in stark contrast to their findings using eligibility criteria of young voters (where frequent voting persists two decades later). 97 It is also worth noting that the initial social pressure outreach studied in the experiments analyzed by Coppock and Green (2016) had some level of novelty and may have triggered awareness at the Attention Stage in a way that temporarily broke a habit of non-voting, by both providing a different contextual cue and changing the expected social costs of continuing to abstain.

One additional strategy used in the past to overcome this exclusion restriction is the use of unexpected shocks like rainfall (Fujiwara, Meng and Vogl, 2016), but as the authors themselves note, it is still very possible that people will still commit a misattribution error and allow an obviously unrelated event color their perceptions of future events. 98 This rain study is also limited by the fact that the unit of observation was county-level data and as the authors note their habit effects may be amplified by social interactions among voters. Indeed, it is worth noting that anticipated social costs of abstention correlate strongly with local turnout rates (Gerber et al., 2016), so a lower-than-anticipated turnout in a county after

---

97 Of course, a very plausible alternative explanation is that this difference in voting persistence may simply be an artifact of two systematically different types of voters.

98 See Achen and Bartels (2017) for prominent examples of voters making misattribution errors.
heavy rain may be communicated by peers and the media, thereby causing an adjustment of
the social value of voting in the Appraisal Stage of the PMBC model.

The data used in this paper does not necessarily circumvent all potential violations of
the exclusion restriction, but it does provide a different perspective in that the exogenous
shock is not solely focused on increasing upstream turnout (as with social pressure nudges),
but also cultivating a habit in the same way public health interventions attempt to make
health behaviors automatic for individuals (e.g., Lally et al 2010). How the cost-benefit
equation of voting would change from this intervention is less clear aside from potentially
reducing the transaction cost of voting.100

4.7: Data

I partnered with a large-scale US labor organization that continually runs door-to-
doctor outreach programs with millions of working class Americans. Their outreach strategy
focuses on building a relationship with voters through recurring contact via in-person
conversations; importantly, both individuals in the control and treatment groups get
subsequent in-person contact.101 Aside from recurring non-election contact ranging from
attempts to increase health insurance uptake to issue-based advocacy, the organization also
does voter mobilization during election years. Election year contact consists of freeform
conversations at the door with several guiding questions, a Get-Out-the-Vote (GOTV)

99 Or reducing turnout, as with rain on an Election Day.

100 A reduction of the transaction cost of voting could be construed as the informational benefit of voting, but
this endogeneity cannot be completely eliminated in any study of in-person voting (for a more detailed
discussion of this mechanism see Meredith, 2009). It is also possible that talking to someone in-person about
an upcoming election may change the social cost or social benefit of voting, but that effect is likely to be
smaller than that of social pressure nudges, because the canvassers’ script did not reference whether or not
someone votes is public record—nor did it instruct canvassers to ask about participants’ prior turnout.

101 Although they emphasize door-to-door contact, they also employ direct mail, phone and digital outreach.
component, and a persuasion component regarding some issue and/or candidate. The
GOTV component is rooted in behavioral psychology research on encouraging habit-
formation. Particularly, the script has the canvasser ask about the voter’s plan to vote (i.e.,
“When do you normally go- on your way to work? When you get off? Just before dinner?”),
which has proven to be an effective GOTV intervention even in high salience election
contexts (Nickerson and Rogers, 2010). The script also includes a component encouraging
voters to talk to their friends and family about the election. (A de-identified sample script
can be found in the Supplementary Materials.)

This treatment is distinct from prior analyses of downstream turnout effects for
three reasons. First, the treatments are specifically tailored around building a voting-habit. Implementation intentions (commonly known as “plan-making nudges”) have proven to have medium to large effects on translating intentions to actions (Gollwitzer and Sheeran, 2008). This component is therefore designed to affect the formation of a cognitive shortcut unlike many of the prior field experiments in the voting-as-a-habit literature. That said, it should be noted that the component of the script that encourages participants to talk to their friends and family about the election could conceivably alter one’s social cost/benefit of voting in the Appraisal Stage. Still, because it does not explicitly reference that whether or not someone votes is public record, it is less likely to affect the cost-benefit equation as

---

102 Additionally, door-to-door canvassing has proven to have the largest persistence effects in prior research (Garcia Bedolla and Michelson, 2012; 188).

103 The impacts on the cost-benefit equation in the Appraisal Stage may still occur. It is possible that the plan-making component may lower the transactional cost of voting by providing implicit information about how much time voting takes, but it should be noted that the vast majority of the individuals in our sample have voted before and already have some information about the voting process. It is therefore less likely that their expected transactional costs of voting will change meaningfully as a result of the conversation. (This would not be the case if the sample was predominantly new voters.)
compared to social pressure GOTV nudges. Social pressure nudges, in contrast, may illustrate that the cost of not voting is higher than expected. An encouragement to simply talk about the election is only likely to affect the social cost of not voting in very indirect ways and is therefore much less likely to affect the cost-benefit calculation than the social pressure intervention. Additionally, recurring contact by the same organization and encouragement to talk about the election with friends and family attempts to keep the same contextual cues for the Attention Stage in future elections, so that the Appraisal stage can more easily be sidestepped. In summary, one key goal of this program was to affect long-term voting behavior of individuals; this stands in contrast to most field experiments in the voting-as-a-habit literature, which were designed to maximize short-term (i.e., upstream) voting behavior.

The partnering organization selected a sample of individuals to be included in their outreach program based on organizational election goals and logistical availability (e.g., the geography had to be dense enough to warrant the cost of door-to-door canvassing). They randomized a percentage of their program sample to an uncontacted control group in nine separate large-N field experiments. The experiments were conducted in 8 different states and included five different election contexts, with a total sample of over 1.8 million households. (A summary of the nine experiments can be found in Table 4.1.) Turnout

---

104 The recurring contact is made without knowledge of the treatment assignment in any given election.

105 The sample is balanced across treatment and control conditions in terms of household income, age (and age^2), sex, race, political party, and whether they voted in the 2012, 2010, and 2008 general elections (chi-squared p-values for each experiment do not exceed 0.05).
among all targets was measured in subsequent elections. Randomization was conducted on the individual level.\textsuperscript{106}

\textsuperscript{106} Some precincts were not canvassed due to programmatic and/or logistical reasons after randomization but before treatment implementation. Since treatment and control units were balanced within precinct, we can exclude these units from the analysis without biasing our results.
Table 4.1: Overview of all experiments

<table>
<thead>
<tr>
<th>Election</th>
<th>Control</th>
<th>Treatment</th>
<th>Total</th>
<th>What's on the ballot?</th>
<th>Population Turnout</th>
<th>Sample Turnout</th>
</tr>
</thead>
<tbody>
<tr>
<td>2014 Mid-Term Election (Iowa)</td>
<td>29,054</td>
<td>28,937</td>
<td>57,991</td>
<td>US Senate, US Congress, State-level offices, 3 ballot initiatives</td>
<td>53%</td>
<td>0.7</td>
</tr>
<tr>
<td>%</td>
<td>50.1</td>
<td>49.9</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014 Mid-Term Election (Illinois)</td>
<td>267,630</td>
<td>268,096</td>
<td>535,726</td>
<td>US Senate, US Congress, State-level offices, 2 ballot initiatives</td>
<td>49%</td>
<td>0.76</td>
</tr>
<tr>
<td>%</td>
<td>49.96</td>
<td>50.04</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014 Mid-Term Election (Michigan)</td>
<td>224,905</td>
<td>224,828</td>
<td>449,733</td>
<td>US Senate, US Congress, State-level offices, 3 ballot initiatives</td>
<td>42%</td>
<td>74.00%</td>
</tr>
<tr>
<td>%</td>
<td>50.01</td>
<td>49.99</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014 Mid-Term Election (New York)</td>
<td>75,679</td>
<td>176,579</td>
<td>252,258</td>
<td>US Congress, State-level Offices, 3 ballot initiatives</td>
<td>33%</td>
<td>0.49</td>
</tr>
<tr>
<td>%</td>
<td>30</td>
<td>70</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014 Mid-Term Election (Washington)</td>
<td>24,004</td>
<td>95,765</td>
<td>119,769</td>
<td>US Congress, 4 ballot initiatives</td>
<td>54%</td>
<td>63.00%</td>
</tr>
<tr>
<td>%</td>
<td>20.04</td>
<td>79.96</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2015 Pennsylvania Supreme Court General Election</td>
<td>16,494</td>
<td>32,784</td>
<td>49,278</td>
<td>3 Supreme Court Seats</td>
<td>27%</td>
<td>60.00%</td>
</tr>
<tr>
<td>%</td>
<td>33.47</td>
<td>66.53</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2015 Philadelphia Mayoral Democratic Primary Election</td>
<td>52,291</td>
<td>153,690</td>
<td>205,981</td>
<td>Democratic Mayoral Primary</td>
<td>26%</td>
<td>57.00%</td>
</tr>
<tr>
<td>%</td>
<td>25.39</td>
<td>74.61</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2016 Presidential Election (North Carolina)</td>
<td>14,544</td>
<td>122,488</td>
<td>137,032</td>
<td>Presidential Election battleground, US Senate, US Congress, State-level offices</td>
<td>69%</td>
<td>0.68</td>
</tr>
<tr>
<td>%</td>
<td>10.47</td>
<td>89.53</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2017 Gubernatorial Election (Virginia)</td>
<td>7,114</td>
<td>60,857</td>
<td>67,971</td>
<td>State-level offices</td>
<td>48%</td>
<td>75%</td>
</tr>
<tr>
<td>%</td>
<td>10.47</td>
<td>89.53</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
As seen in Tables 4.2-4.3, the demographics of the sample in each experiment differed substantially. These differences are partially an artifact of geography. (e.g., individuals randomized in the Philadelphia experiment are predominantly black, because Philadelphia is a black majority city). But a major part of these differences is due to strategic considerations: namely, the organization is tasked with engaging with working class voters and so their targets will not be representative of the wider electorate. Furthermore, the organization’s sampling criteria reflected their expectations as to people who are likely to be receptive to their GOTV outreach (which explains why such a high proportion of individuals across all experiments voted in the 2012 general election). To account for demographic differences in the samples, I control for all available observable covariates.\textsuperscript{107} Still, no amount of control covariates can capture certain intrinsic differences between populations and so one drawback of some of the subsequent subgroup analyses is that they may be driven by unobservable differences in the sampling criteria.

\textsuperscript{107} I do not use sampling weights as weighting to a particular population will inevitably reflect arbitrary choices by the researcher (for a more thorough discussion of sampling weights, see Gelman (2007)).
Table 4.2: Summary statistics (demographics)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>18-29</td>
<td>6%</td>
<td>10%</td>
<td>3%</td>
<td>3%</td>
<td>8%</td>
<td>6%</td>
<td>7%</td>
<td>1%</td>
<td>27%</td>
<td>8%</td>
<td></td>
</tr>
<tr>
<td>30-49</td>
<td>28%</td>
<td>35%</td>
<td>21%</td>
<td>32%</td>
<td>26%</td>
<td>27%</td>
<td>34%</td>
<td>29%</td>
<td>40%</td>
<td>25%</td>
<td></td>
</tr>
<tr>
<td>50-64</td>
<td>36%</td>
<td>30%</td>
<td>37%</td>
<td>43%</td>
<td>34%</td>
<td>33%</td>
<td>34%</td>
<td>35%</td>
<td>21%</td>
<td>29%</td>
<td></td>
</tr>
<tr>
<td>65+</td>
<td>30%</td>
<td>25%</td>
<td>38%</td>
<td>22%</td>
<td>32%</td>
<td>33%</td>
<td>25%</td>
<td>35%</td>
<td>12%</td>
<td>38%</td>
<td></td>
</tr>
<tr>
<td>Gender</td>
<td>% Female</td>
<td>54%</td>
<td>60%</td>
<td>53%</td>
<td>48%</td>
<td>54%</td>
<td>53%</td>
<td>60%</td>
<td>60%</td>
<td>56%</td>
<td>59%</td>
</tr>
<tr>
<td>Race</td>
<td>White</td>
<td>72%</td>
<td>0%</td>
<td>83%</td>
<td>83%</td>
<td>91%</td>
<td>86%</td>
<td>41%</td>
<td>36%</td>
<td>34%</td>
<td>81%</td>
</tr>
<tr>
<td></td>
<td>Black</td>
<td>17%</td>
<td>0%</td>
<td>8%</td>
<td>11%</td>
<td>2%</td>
<td>3%</td>
<td>52%</td>
<td>57%</td>
<td>51%</td>
<td>18%</td>
</tr>
<tr>
<td></td>
<td>Hispanic</td>
<td>4%</td>
<td>0%</td>
<td>6%</td>
<td>1%</td>
<td>5%</td>
<td>3%</td>
<td>5%</td>
<td>5%</td>
<td>6%</td>
<td>0%</td>
</tr>
<tr>
<td>Partisanship</td>
<td>Democrat</td>
<td>24%</td>
<td>56%</td>
<td>0%</td>
<td>0%</td>
<td>30%</td>
<td>0%</td>
<td>98%</td>
<td>98%</td>
<td>67%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>Republican</td>
<td>6%</td>
<td>5%</td>
<td>0%</td>
<td>0%</td>
<td>39%</td>
<td>0%</td>
<td>2%</td>
<td>2%</td>
<td>5%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>Party Unknown</td>
<td>66%</td>
<td>0%</td>
<td>100%</td>
<td>100%</td>
<td>23%</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Median Household Income</td>
<td>$50,000 - $75,000</td>
<td>$50,000</td>
<td>$50,000</td>
<td>$50,000</td>
<td>$50,000</td>
<td>$50,000</td>
<td>$30,000</td>
<td>$30,000</td>
<td>$30,000</td>
<td>$30,000</td>
<td>$50,000</td>
</tr>
<tr>
<td>% Married</td>
<td>47%</td>
<td>51%</td>
<td>66%</td>
<td>66%</td>
<td>66%</td>
<td>66%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>28%</td>
<td>0%</td>
</tr>
<tr>
<td>Total N</td>
<td>1,875,739</td>
<td>57,991</td>
<td>535,726</td>
<td>449,733</td>
<td>252,258</td>
<td>119,769</td>
<td>49,278</td>
<td>205,981</td>
<td>137,032</td>
<td>67,971</td>
<td></td>
</tr>
</tbody>
</table>
Table 4.3: Summary statistics (vote history)

<table>
<thead>
<tr>
<th>% Voted in 2012 General</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Whole Sample</td>
<td>90%</td>
<td>99%</td>
<td>93%</td>
<td>98%</td>
<td>83%</td>
<td>92%</td>
<td>92%</td>
<td>90%</td>
<td>60%</td>
<td>91%</td>
<td></td>
</tr>
<tr>
<td>IA 2014</td>
<td>73%</td>
<td>76%</td>
<td>87%</td>
<td>79%</td>
<td>56%</td>
<td>73%</td>
<td>67%</td>
<td>74%</td>
<td>27%</td>
<td>62%</td>
<td></td>
</tr>
<tr>
<td>IL 2014</td>
<td>85%</td>
<td>85%</td>
<td>89%</td>
<td>92%</td>
<td>81%</td>
<td>80%</td>
<td>85%</td>
<td>91%</td>
<td>52%</td>
<td>84%</td>
<td></td>
</tr>
<tr>
<td>MI 2014</td>
<td>85%</td>
<td>85%</td>
<td>89%</td>
<td>92%</td>
<td>81%</td>
<td>80%</td>
<td>85%</td>
<td>91%</td>
<td>52%</td>
<td>84%</td>
<td></td>
</tr>
<tr>
<td>NY 2014</td>
<td>85%</td>
<td>85%</td>
<td>89%</td>
<td>92%</td>
<td>81%</td>
<td>80%</td>
<td>85%</td>
<td>91%</td>
<td>52%</td>
<td>84%</td>
<td></td>
</tr>
<tr>
<td>WA 2014</td>
<td>85%</td>
<td>85%</td>
<td>89%</td>
<td>92%</td>
<td>81%</td>
<td>80%</td>
<td>85%</td>
<td>91%</td>
<td>52%</td>
<td>84%</td>
<td></td>
</tr>
<tr>
<td>PA 2015</td>
<td>85%</td>
<td>85%</td>
<td>89%</td>
<td>92%</td>
<td>81%</td>
<td>80%</td>
<td>85%</td>
<td>91%</td>
<td>52%</td>
<td>84%</td>
<td></td>
</tr>
<tr>
<td>Philly 2015</td>
<td>85%</td>
<td>85%</td>
<td>89%</td>
<td>92%</td>
<td>81%</td>
<td>80%</td>
<td>85%</td>
<td>91%</td>
<td>52%</td>
<td>84%</td>
<td></td>
</tr>
<tr>
<td>NC 2016</td>
<td>85%</td>
<td>85%</td>
<td>89%</td>
<td>92%</td>
<td>81%</td>
<td>80%</td>
<td>85%</td>
<td>91%</td>
<td>52%</td>
<td>84%</td>
<td></td>
</tr>
<tr>
<td>VA 2017</td>
<td>85%</td>
<td>85%</td>
<td>89%</td>
<td>92%</td>
<td>81%</td>
<td>80%</td>
<td>85%</td>
<td>91%</td>
<td>52%</td>
<td>84%</td>
<td></td>
</tr>
<tr>
<td>% Voted in 2008 General</td>
<td>85%</td>
<td>85%</td>
<td>89%</td>
<td>92%</td>
<td>81%</td>
<td>80%</td>
<td>85%</td>
<td>91%</td>
<td>52%</td>
<td>84%</td>
<td></td>
</tr>
<tr>
<td>% Voted in 2010 General</td>
<td>73%</td>
<td>76%</td>
<td>87%</td>
<td>79%</td>
<td>56%</td>
<td>73%</td>
<td>67%</td>
<td>74%</td>
<td>27%</td>
<td>62%</td>
<td></td>
</tr>
<tr>
<td>% Voted in 2012 General</td>
<td>90%</td>
<td>99%</td>
<td>93%</td>
<td>98%</td>
<td>83%</td>
<td>92%</td>
<td>92%</td>
<td>90%</td>
<td>60%</td>
<td>91%</td>
<td></td>
</tr>
<tr>
<td>Total N</td>
<td>1,875,739</td>
<td>57,991</td>
<td>535,726</td>
<td>449,733</td>
<td>252,258</td>
<td>119,769</td>
<td>49,278</td>
<td>205,981</td>
<td>137,032</td>
<td>67,971</td>
<td></td>
</tr>
</tbody>
</table>
4.7.1: Predictions

This data gives us the opportunity to test concrete predictions borne out of the PMBC and the voting-as-a-habit literature, while avoiding the file-drawer effect of conventional meta-analyses and literature overviews. I explicate each prediction in turn.

**Prediction 1:** Voters induced to vote in one election type (e.g., presidential) will exhibit stronger vote consuetude effects in future elections of the same type. To bypass the Appraisal Stage, the contextual cues in the Attention Stage need to be similar to previous instances where the outcome from the Situation Stage was positive. As such, more similar contextual cues in an election make it more likely that an individual will bypass the Appraisal Stage as a direct result of prior behavior.

**Prediction 2:** Voters induced to vote in a low-salience election will not exhibit stronger vote consuetude effects. Contextual cues of a lower salience election are likely different from a high salience election. Higher salience elections have greater social and media coverage. As such, the context cues in the higher salience election may call attention to the election in a different way, which would make it less likely for an individual to take a cognitive shortcut to turn out. Rather, the novel cues are more likely to trigger a re-appraisal of the cost-benefit equation.

**Prediction 3:** Voting in an election where one’s chosen candidate won should be positively associated with voting consuetude formation. The PMBC model stipulates that positive reinforcement is key to strengthening the heuristic shortcut past the Appraisal Stage. As such, casting a vote in an election where your preferred candidate wins should provide a larger “warm glow” than individuals who voted and their candidate lost. Since a larger warm glow term makes it more likely that the outcome of the cost-benefit analysis in the Appraisal
Stage is positive, the individual would thus get some positive reinforcement, which should make subsequent voting in an election with similar context cues more likely.

**Prediction 4**: *Household income may be associated with voting consuetude formation.* This final prediction is dependent on a number of assumptions that have mixed support in the existing literature. Fujiwara, Meng and Vogl (2016) noted that the cost of voting is relatively higher for individuals with lower household incomes. On the other hand, the opportunity cost of voting for higher income individuals may be higher and thus discourage some higher income individuals from voting. A meta-analysis of 40 published studies found that income was positively correlated with likelihood to turn out, but the relationship failed to be significant in nearly half of the studies included in the meta-analysis (Smets & van Ham 2013). As such, this prediction rests on precarious theoretical ground and the empirical results from this prediction should be viewed with a healthy dose of skepticism.

That said, the proposed mechanism is as follows. There is a documented association between lower levels of income and higher levels of geographic mobility (e.g., Purcell, 2020), which would imply that contextual cues in different geographies are likely to be different and hence detrimental to the formation of heuristic shortcuts. This is consistent with Coppock & Green’s (2016) conclusions that movers were less likely to exhibit voting consuetude. And so, the PMBC framework implies that lower income individuals would be less likely to exhibit voting consuetude.

**4.8: Results**

This section is organized as follows. First, I analyze the upstream effects of the experiments in this sample. I then explicate my analysis strategy for voting consuetude and
analyze the main effects. Finally, I evaluate predictions as to which circumstances are more (or less) conducive to voting consuetude.

4.8.1: Upstream Effects

As a first step, I analyze all the field experiments together in an OLS regression (with experiment fixed effects and robust standard errors) to measure the impact of the treatment on upstream elections. Overall, the organization’s GOTV outreach increased turnout by 0.3 percentage points (Intent-to-Treat, se = 0.0007, p<0.001). It is standard practice (e.g., Gerber and Green 2012) to take into account that not all households were successfully contacted and evaluate the Complier Average Causal Effect (CACE), where the complier is the individual who was successfully contacted by a canvasser.108 I find that, at the door, the organization’s GOTV canvass had a 1.8 percentage point impact on voter turnout (CACE, se = 0.0038, p<0.001). This overall effect is nearly identical to the most recent meta-analysis of door-to-door canvassing (Green & Gerber, 2019), which similarly finds a CACE of 1.8.109 This is all to say that the upstream impact of the doorknock treatment in this paper is typical of door-to-door mobilization efforts.

4.8.2: Analysis Strategy

I use the analysis strategy leveraged by Coppock and Green (2016), where $V_1$ is defined as voting in the upstream election and $V_2$ is voting in a downstream election, $Z$ is an indicator denoting whether or not the individual was assigned to receive a doorknock

---

108 This does not necessarily mean that the entire script was delivered. “Successfully contacted” denotes that the individual on the canvass list opened the door and affirmatively identified themselves.

109 The meta-analysis decomposes treatment effects by turnout in the control group. Among the studies that had a control group turnout rate between 50% and 70% (the control group in my data had a turnout rate of 67.6%), the CACE was 1.8 with a 95% confidence interval of [0.4, 3.3] (Green & Gerber 2019, p. 211).
GOTV treatment at time 1. The main estimand of interest is a (different) CACE—the effect of voting in an upstream election on downstream voting among those who vote because they receive the GOTV doorknock. The estimator is thus

\[
\text{CACE} = \frac{\mathbb{E}[V_{2i}|Z_i=1] - \mathbb{E}[V_{2i}|Z_i=0]}{\mathbb{E}[V_{1i}|Z_i=1] - \mathbb{E}[V_{1i}|Z_i=0]}
\]

for every household \( i \) in the experiment. This is estimated via two-stage least squares as is standard practice (Angrist, Imbens, & Rubin, 1996).

To check for weak instruments, I examine the impact of each treatment on upstream elections by experiment (see Table 4.4). Though some experiments were underpowered on their own, most of the outreach programs had upstream treatment effects that were well within range of one another. The only exception is the 2014 canvass program in Washington, which appears to have had a near-perfect zero impact on turnout in the 2014 general election (\( p=0.95 \)). The extremely small magnitude of the upstream coefficient is likely to blow up our CACE estimates in this experiment. As such, in all subsequent analyses, the Washington data is omitted. Please note that including the Washington data in all subsequent analyses does not substantially change the reported results since it accounted for just \( \sim 6\% \) of the entire sample.

---

110 As in Coppock and Green (2016), we must assume non-interference, which may be more plausible in our case, as the intervention does not publicize the vote history of neighbors, which may be a topic of conversation in close-knit communities.

111 I report all downstream CACEs for all experiments in the Supplementary Materials and find that this is indeed the case.
Table 4.4: Upstream effects by each experiment

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Intent-to-treat Effect on Upstream Election</th>
<th>Treatment-on-Treated Effect on Upstream Election</th>
</tr>
</thead>
<tbody>
<tr>
<td>2014 LA</td>
<td>0.0140</td>
<td>0.0666</td>
</tr>
<tr>
<td>n=57,991</td>
<td>(0.0034)</td>
<td>(0.0164)</td>
</tr>
<tr>
<td>2014 IL</td>
<td>0.0016</td>
<td>0.0101</td>
</tr>
<tr>
<td>n=535,726</td>
<td>(0.0010)</td>
<td>(0.0063)</td>
</tr>
<tr>
<td>2014 MI</td>
<td>0.0022</td>
<td>0.0112</td>
</tr>
<tr>
<td>n=449,733</td>
<td>(0.0012)</td>
<td>(0.0061)</td>
</tr>
<tr>
<td>2014 NY</td>
<td>0.0019</td>
<td>0.0139</td>
</tr>
<tr>
<td>n=252,258</td>
<td>(0.0018)</td>
<td>(0.0136)</td>
</tr>
<tr>
<td>2014 WA</td>
<td>0.0002</td>
<td>0.0006</td>
</tr>
<tr>
<td>n=119,769</td>
<td>(0.0030)</td>
<td>(0.0104)</td>
</tr>
<tr>
<td>2015 PA SUPREME COURT</td>
<td>0.0116</td>
<td>0.0652</td>
</tr>
<tr>
<td>n=49,278</td>
<td>(0.0045)</td>
<td>(0.0254)</td>
</tr>
<tr>
<td>2015 PHILLY MAYORAL</td>
<td>0.0078</td>
<td>0.0249</td>
</tr>
<tr>
<td>n=205,981</td>
<td>(0.0023)</td>
<td>(0.0074)</td>
</tr>
<tr>
<td>2016 NC</td>
<td>0.0024</td>
<td>0.0532</td>
</tr>
<tr>
<td>n=137,032</td>
<td>(0.0036)</td>
<td>(0.0817)</td>
</tr>
<tr>
<td>2017 VA</td>
<td>0.0074</td>
<td>0.0497</td>
</tr>
<tr>
<td>n=67,971</td>
<td>(0.0053)</td>
<td>(0.0354)</td>
</tr>
</tbody>
</table>
4.8.3: Overall Downstream Effects

Since the initial canvasses were conducted anywhere from 2014 to 2017, the only downstream election for which we can use all of the data is the 2018 general election.\footnote{Due to administrative issues, the 2018 primary turnout was never appended to my data.} When I regress data from all experiments (with experiment fixed effects and robust standard errors) in a 2SLS regression with voting in the 2018 general election as the instrumented variable, I estimate a CACE of 0.57 (p=0.004).\footnote{It is possible to analyze downstream CACEs for 2016 primary and general elections for a smaller subset of my data. I do this in a subsequent analysis below.} This “habit” effect is somewhat lower than Fujiwara, Meng, and Vogl (2016)’s estimate of 0.9, though the 95% confidence interval [0.18, 0.96] for my data does include 0.9. Still, the somewhat higher CACE in Fujiwara, Meng, and Vogl (2016) is consistent with their explanation that their effect is possibly inflated by social interactions. My CACE estimate is somewhat higher than the ~4-year downstream CACEs reported in Coppock and Green (2016), which range from 0.13 to non-significant (and directionally negative). This somewhat more muted voting consuetude comports well with the PMBC framework. Namely, the Coppock an Green (2016) experiments all had treatments that focused on impacting an individual’s cost-benefit equation (whether it was emphasizing civic duty, using social pressure, or simply informing voters that they were being observed), which, if successful, should temporarily impact the cost-benefit equation among compliers. Though it is also possible that these treatments did create voting heuristics that simply faded due to some mixture of a lack of positive reinforcement (e.g., bad experiences at the polls) or a lack of similar contextual cues. However, the persistence of
consuetude across different election types suggests that it is more likely that the cost-benefit equation was impacted.

4.8.4: Evaluating Predictions from Prior Research

The most important contribution of this study is that the variety of election contexts and the large sample-size allows us to independently test predictions as to the circumstances in which voting consuetude is more (or less) likely to occur. This allows us to better understand the mechanism behind voting consuetude. I address each prediction in turn below. Table 4.5 summarizes my predictions, my empirical findings, and the empirical findings of previous studies that looked at a similar population of voters (i.e., not new voters).
Table 4.5: Summary of predictions & results across multiple studies

<table>
<thead>
<tr>
<th>Study</th>
<th>Ternovski 2021 N=~1.8M 9 experiments</th>
<th>Coppock &amp; Green 2016 N=~1.2M 3 experiments</th>
<th>Fujiwara, Meng &amp; Vogl 2016 N=~50k 1 natural experiment</th>
<th>Hill &amp; Kouser 2016 N=~150k 1 experiment</th>
<th>Garcia Bedolla &amp; Michelson 2012 N=~133k 14 experiments</th>
<th>Michelson 2003 N=~3k 1 experiment</th>
<th>Gerber, Green, and Shachar 2003 N=~25k 1 experiment</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>P1:</strong> Voters induced to vote in one election type will exhibit stronger vote consuetude effects in future elections of the same type.</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes (mixed)*</td>
<td>Yes**</td>
<td>No**</td>
<td>Yes**</td>
<td>No**</td>
</tr>
<tr>
<td><strong>P2:</strong> Voters induced to vote in a low-salience election will not exhibit stronger vote consuetude effects.</td>
<td>Yes</td>
<td>No</td>
<td>N/A</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>P3:</strong> Voting in an election where one’s chosen candidate won should be positively associated with voting consuetude formation.</td>
<td>Yes</td>
<td>N/A</td>
<td>No</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>P4:</strong> Household income may be associated with voting consuetude formation.</td>
<td>Yes</td>
<td>N/A</td>
<td>Yes</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

* = One set of studies did yield results that were consistent with my prediction, but a second set of studies did not. The authors believe the second set failed to replicate due to inadequate statistical power. ** = These studies only had data on the effect of voting in one election type upstream on voting in a different type of election downstream. Thus, we are unable to determine if the effects they find are higher for elections of the same type (i.e., we have no counterfactual within study).
**Prediction 1:** Voters induced to vote in one election type (e.g., presidential) will exhibit stronger vote consuetude effects in future elections of the same type. According to PMBC, contextual cues are key to triggering heuristic shortcuts past the appraisal stage. Since different types of elections have different levels of media coverage, voting consuetude should be stronger for elections of the same type.

In my data, there are four cases where the upstream and downstream election is of the same type (all federal mid-term elections). In Table 4.6, I present the initial upstream turnout effect in the first column, the CACE of mid-term turnout on turnout in a downstream presidential primary and general elections in the next two columns, and the CACE of mid-term turnout on a mid-term downstream election in the final column. It does appear that downstream voting persistence among compliers was small to null in the presidential election, but was very strong in the following mid-term election. My data therefore comports strongly to both the PMBC model and Coppock and Green’s (2016) finding.

It is important to point out that the conventional explanation for these differences would focus on ceiling effects. Particularly, if the base rate of voting in the control group is very high, the maximum possible effect size among compliers is therefore much lower than an election with a low turnout rate. The turnout rates in the control group for 2016 general

---

115 As mentioned in the Analysis Strategy subsection, Washington data was excluded due to the lack of initial upstream effects. The inclusion of this data does not meaningfully change the results. I also omit the 2016 North Carolina presidential experiment from this comparison as I do not have data for the 2020 presidential election; that said, the downstream effect in North Carolina on the following mid-term election was not significant (p=0.277), but the magnitude was high (0.69). As such, it is likely this analysis is underpowered and I therefore cannot draw any conclusions from the North Carolina analysis.

116 Note that there is a somewhat higher (but non-significant) CACE in the presidential primary, which may be driven by lower base rate of turnout as predicted in Coppock and Green (2016).
and the 2018 general elections in my sample are not dramatically different (74% in the 2018 midterm versus 86% in the 2016 presidential) and the CACE for the 2018 midterm is over four times larger than the CACE for the 2016 general. As such, I conclude the PMBC explanation is more likely than the conventional base rate explanation.

**Table 4.6: CACEs and election type**

<table>
<thead>
<tr>
<th></th>
<th>2014 General Election (1st Stage)</th>
<th>2016 Presidential Primary Election</th>
<th>2016 Presidential General Election</th>
<th>2018 Mid-term General Election</th>
</tr>
</thead>
<tbody>
<tr>
<td>Midterms Only (IA, IL, MI, NY)</td>
<td>0.0023 (.0007)</td>
<td>0.4246 (.3176)</td>
<td>0.2212 (.2255)</td>
<td>0.959 (.3506)</td>
</tr>
</tbody>
</table>

**Prediction 2:** Voters induced to vote in a low-salience election will not exhibit stronger vote consuetude effects. This prediction is related to Prediction 1. PMBC predicts that context cues will be different in different types of elections and voting in a lower salience election will be a markedly different event (in terms of, for instance, media coverage) as compared to voting in a presidential election.

A common means of defining salience is examining the actual turnout on Election Day (e.g., see Coppock and Green, 2016). In our data, the 2015 PA Supreme Court election and the 2015 Philadelphia Mayoral Primary clearly qualify with turnout well under 30%.117 As seen in Table 4.7 below, despite the fact that the canvass in Pennsylvania and Philadelphia had an upstream treatment effect size that was about 3.4 times as large as that of the rest of the experiments, we find that there were no lasting downstream effects among

---

117 And indeed, local media coverage of these election has called turnout “low” (Holmberg 2015) or even “bad” (Kerkestra 2015).
compliers. In contrast, nearly 95% of participants induced to vote by the initial GOTV contact in the remaining states turned out downstream in the 2018 mid-term election. One alternative explanation to the PMBC interpretation is that the compliers who were successfully turned to vote in the 2015 PA and the 2015 Philadelphia elections were already high-propensity voters and their lack of subsequent voting consuetude is simply that they are already voting in higher salience elections. However, we find that the control groups in the 2015 PA and the 2015 Philadelphia studies turned out at 72.3% in the 2018 mid-term election, while all other individuals in the controls groups of my sample turned out at a rate 73.8% in the 2018 mid-term election. As such, we should not expect such a strong divergence in downstream turnout effects if they are simply driven by a turnout ceiling.

Still, one could construe the NY 2014 general election as low salience, as well. Even though there was a gubernatorial election, Andrew Cuomo was predicted to win handedly (RCP 2014) and turnout was only seven percentage points higher than the Philadelphia Mayoral race. When I include NY as one of the three low salience elections (in the second row of Table 4.7), there is, again, no persistence in the 2018 general election among these low salience experiments, despite an upstream effect that’s nearly twice as large as the upstream effect of all other experiments. The other experiments, in contrast, again have large and statistically significant CACEs in the 2018 general election.

Salience is not only measured by the actual turnout; it can be construed as races that lack higher office seats on the ballot. In the remaining two rows, I try two more combinations of low salience races and, in both cases, find null downstream CACEs. This stands in contrast to Coppock and Green (2016) and Garcia Bedolla and Michelson’s (2012) findings but is consistent with Michelson (2003) and Hill & Kousser (2016).
Table 4.7: Salience and downstream CACEs

<table>
<thead>
<tr>
<th>Event Description</th>
<th>Low Salience (PA, Philly)</th>
<th>High Salience (All others)</th>
<th>2018 Mid-term General Election</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lowest Projected Turnout + No Federal Race on Ballot</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low Salience</td>
<td>0.0085</td>
<td>0.025</td>
<td>0.0085 (0.0021)</td>
</tr>
<tr>
<td>High Salience</td>
<td></td>
<td>0.8829</td>
<td>0.0025 (0.0007)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lowest Projected Turnout</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low Salience</td>
<td>0.0050</td>
<td>0.0028</td>
<td>0.0050 (0.0014)</td>
</tr>
<tr>
<td>High Salience</td>
<td></td>
<td>0.9463</td>
<td>0.0028 (0.0007)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No Federal Seats on Ballot</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low Salience</td>
<td>0.00835</td>
<td>0.024</td>
<td>0.00835 (0.0019)</td>
</tr>
<tr>
<td>High Salience</td>
<td></td>
<td>0.9038</td>
<td>0.0024 (0.0007)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No President/Governor/Senate Race on Ballot</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low Salience</td>
<td>0.0053</td>
<td>0.00792</td>
<td>0.0053 (0.0013)</td>
</tr>
<tr>
<td>High Salience</td>
<td></td>
<td>0.9717</td>
<td>0.00792 (0.0008)</td>
</tr>
</tbody>
</table>

**Prediction 3:** Voting in an election where one’s chosen candidate won should be positively associated with voting consuetude formation. This prediction is borne out of the fact that heuristic-shortcut formation should only occur through positive reinforcement of the behavior. While
it is impossible to know how much positive utility an individual gets out of voting in a particular election, we have do have one approximate metric available—whether an individual’s chosen candidate won. Although, it is not possible to know who a given participant voted for, the targeting criteria of our partner organization is overwhelmingly Democratic. That said, party registration is missing from a large portion of our data. So, for this analysis, I remove registered Republicans and assume individuals with missing registration data prefer the Democratic candidate to win. The degree to which this assumption is violated will determine how accurate this analysis is. With these caveats in mind, I find that participating in an upstream election where the Democratic candidate won was associated with persistent downstream effects among compliers four years later. This was not the case in states where the Republican won (see Table 8). I should emphasize that these results are only based on a comparison of two sets of experiments and there may be other confounding environmental variables that are responsible for this difference in voting consuetude.

**Table 4.8: Race Outcome and CACEs**

<table>
<thead>
<tr>
<th></th>
<th>Upstream Election (1st Stage)</th>
<th>2018 Mid-term General Election</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dem Win (IL 2014, MI 2014)</strong></td>
<td>0.0019</td>
<td>1.3969</td>
</tr>
<tr>
<td></td>
<td>(.0008)</td>
<td>(.6173)</td>
</tr>
<tr>
<td><strong>Dem Loss (IA 2014, NC 2016)</strong></td>
<td>0.0090</td>
<td>0.2013</td>
</tr>
<tr>
<td></td>
<td>(.0026)</td>
<td>(.2856)</td>
</tr>
</tbody>
</table>

---

118 Since the election context is different in the 2017 Virginia experiment, I exclude it from this comparison. However, adding the VA data results in a highly similar CACE of 1.31 (p=0.02).
**Prediction 4:** Household income may be associated with voting consuetude formation. When I look at the initial upstream effect among participants with lower than the mean income level in our sample ($50,000 - $75,000), the upstream treatment effect is actually higher than the upstream treatment effect for participants who have a higher than mean level of income (0.43 percentage points [p<.01] vs. 0.20 percentage points [p=.04]). In other words, the GOTV outreach in this study appears to be more effective for lower income individuals in upstream elections. But the downstream CACE for higher income households is twice as large as for lower income households (1.1 [p=.08] vs. 0.45 [p=.02]). In short, these findings offer some support for the PMBC prediction that voting consuetude is less likely to occur for lower income voters.

### 4.9: Discussion and Conclusion

My results indicate that the PMBC theoretical framework may provide additional insight into why voting consuetude forms in some contexts but not others. Its predictions appear to validate in a novel large-scale dataset, but since the predictions were not preregistered before the data was collected, it is imperative to replicate these results in future studies and assess some of the alternative explanations that might yield the same results. It is also worth noting that even though the experiments were conducted in different election contexts in different states, the types of voters included in these samples are nevertheless not nationally representative and reflect the labor organization’s strategic targeting criteria (i.e., working-class, Democratic-leaning voters). As such, it is unclear whether these results generalize to a broader population. Additionally, though I assert that the treatment did not...

---

119 The income variable is modelled by the data vendor, Catalist, by appending commercial consumer data to the voter file.
include any social pressure interventions (which might credibly change the cost-benefit equation in the short-term), the conversations were freeform and we have no way of knowing if there were large numbers of canvassers who went off-script. This possible limitation is, however, unlikely, as the labor organization employs paid canvassers who receive extensive training.

One possible alternative explanation to voting consuetude is raised by what Rogers and Frey (2014) termed “rip currents,” where compliance to an upstream GOTV intervention leads to subsequent attention from campaigns and non-profits. In other words, this alternative explanation claims that it’s not that there is a permanent change to an individual’s cost-benefit equation or the act of voting creates a self-reinforcing intrinsic impetus to vote; rather, the rip currents hypothesis claims that the people who are successfully nudged to the polls by a campaign are subsequently targeted by more campaigns and non-profits with nudges that are similarly successful in inducing these individuals to vote in downstream elections.

However, the empirical evidence suggests that this is unlikely to be the main driver behind voter consuetude. Two studies addressed this theory and found little evidence in support of this hypothesis. First, Coppock and Green (2016) find that voting consuetude isn’t more pronounced in battleground states where one would except stronger “rip currents.” In other words, since battleground states attract more campaign activity, those voters who are successfully mobilized in an upstream election should receive more campaign attention under the rip current hypothesis.120 However, one drawback of this analysis is that

---

120 Coppock and Green (2016) also provide an overview of other studies that looked at subsequent campaign contact across upstream treatment conditions; Dinas (2012) and Green, McGrath, & Aronow (2013) found very modest effects of treatment on subsequent campaign contact.
more campaign activity may also mean more people are targeted—even those who were not successfully mobilized\textsuperscript{121}—which may dilute any rip current effects. The second piece of evidence is the analysis in Rogers et al. (2017), where the authors explicitly analyzed whether campaigns and non-profits were more likely to engage in downstream contact of participants who were in an upstream treatment group (as compared to participants in control). The initial field experiment included \textasciitilde664,000 Democratic-leaning subjects randomized to GOTV outreach before the 2011 Senate Recall Election in Wisconsin. In 2016, these subjects were subsequently matched to two databases of campaign contact information.\textsuperscript{122} While Rogers et al. (2017) did find some evidence of increased levels of downstream outreach of individuals in the 2011 treatment group, the differences were modest and the authors concluded that “the downstream turnout effects… [that they found] cannot plausibly be attributed to the treatment and control groups' differential exposure to mobilization activity.” (p. 92).\textsuperscript{123}

Furthermore, an increase in subsequent campaign contact still comports to the PMBC theoretical framework in that repeat outreach may mean that individuals are receiving similar context cues in future elections. The key question is whether or not the subsequent campaign contact affects the Appraisal Stage or simply calls attention to an upcoming

\textsuperscript{121} One common campaign strategy is discussed in Arceneaux & Nickerson’s (2008) meta-analysis: “[i]n high-salience elections, campaigns target unlikely voters out of the belief that everyone else is going to vote without their encouragement, whereas in low-salience elections they assume the opposite and focus on those voters who have reliably voted in the past” (p. 5).

\textsuperscript{122} The first database was maintained by the Obama for America campaign (which ran Barak Obama’s campaign in 2008 and 2012) and the second, by Catalist, a Democratic-leaning data clearinghouse that tracked voter contact data of several large Democratic-leaning non-profits (Rogers et al. 2017).

\textsuperscript{123} The only sizable increase came in the form of direct mail (8.1 percentage points more in the treatment group), which is unlikely to explain the entirety of consuetude effects. A meta-analysis of direct mail outreach finds a 0.5 percentage point impact; even the most effective intervention (social pressure) has, on average, only a 2 percentage point impact on turnout (Green & Gerber 2019; p. 214).
election. Future research should attempt to disentangle these effects by comparing outreach that targets variables in the cost-benefit equation versus outreach that consists of repeating context cues. For instance, one possible experimental design could test continual informational treatments (e.g., identical text messages reminding an individual of an upcoming election day every election) versus the impact of a one-off, upstream social pressure treatment on downstream turnout. PMBC would predict that while the information treatment should have a lower upstream effect, compliers in the informational treatment should develop stronger voter consuetude years later (as compared to compliers in the social pressure condition).

Another related alternative explanation is that voting starts a self-reinforcing process (e.g., Rogers & Frey 2014). This explanation is related to the informational boluses one gets from voting, which may include iteratively reducing transaction costs (e.g., learning the best time to go to one’s local polling place), inflating the warm glow of voting or the social benefits of voting. Under this explanation, the economic model is sufficient to explain voting consuetude without the need for the PMBC theoretical framework. However, it is difficult to reconcile this mechanism with, for instance, Coppock and Green’s (2016) finding that downstream treatment effects of social pressure mailings eventually faded.124 If voting consuetude is always driven by a self-reinforcing adjustment of costs, we would expect that the specifics of the initial motivation for voting shouldn’t affect the longevity of effects.

124 Previously, Davenport et al. (2010) found similar decays in downstream effects from social pressure interventions. (There is some overlap in data used in Coppock and Green (2016) and Davenport et al. (2010); to be clear, in this footnote, I refer to studies included in Davenport et al. (2010) but not included in Coppock and Green (2016).)
There may still be other explanations that do not comport with the PMBC model and in no way, should my empirical assessment be viewed as conclusive evidence that voting consuetude is explained solely by the PMBC theoretical framework. The goal of this paper is to present a new theoretical framework with empirical evidence illustrating this framework’s value. Future studies should directly and independently test the predictions that come out of my application of PMBC to turnout persistence. Specifically, I predict to increase the chances of developing voting consuetude (particularly among older voters), the following conditions should hold. First, the contextual cues in one election should be as similar as possible in future elections. This implies that organizations tasked with increasing turnout may want to attempt using the same prompt to remind voters that it’s Election Day from one election to the next. Second, it is important that voters successfully persuaded to turn out to the polls have a net positive experience voting. This makes it more likely that a voter will begin to skip the Appraisal Stage and adopt a cognitive shortcut in response to a contextual cue. The implication is the voters who did not have a positive experience voting (e.g., long lines, their preferred candidate lost) may benefit from receiving a different contextual cue next election to stop the development of a non-voting habit. And finally, for voting to become truly habitual, repetition is key. This means that in certain contexts and with frequent elections, a voting habit in the strict-psychological sense maybe possible. For instance, Solvack and Vassil (2018) find promising persistence effects in online voting in Estonia.
4.10: References


4.11: Supplementary Materials

4.11.1: Additional Tables and Figures

**Figure 4.2:** Average treatment effects for upstream and downstream elections by experiment
Table 4.9: Upstream intent-to-treat effect and CACEs for downstream elections by experiment

<table>
<thead>
<tr>
<th>Experiment</th>
<th>1st Stage (Upstream Election)</th>
<th>2016 Primary CACE</th>
<th>2016 General CACE</th>
<th>2018 General CACE</th>
</tr>
</thead>
<tbody>
<tr>
<td>2014 IA</td>
<td>0.0140 (0.0034)</td>
<td>-0.1727 (0.2136)</td>
<td>0.0701 (0.1963)</td>
<td>0.2225 (0.2333)</td>
</tr>
<tr>
<td>n=57,991</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014 IL</td>
<td>0.0016 (0.0010)</td>
<td>1.2897 (0.9389)</td>
<td>0.6913 (0.5664)</td>
<td>1.7542 (1.1250)</td>
</tr>
<tr>
<td>n=535,726</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014 MI</td>
<td>0.0022 (0.0012)</td>
<td>0.4255 (0.5964)</td>
<td>-0.2773 (0.4953)</td>
<td>1.0485 (0.6810)</td>
</tr>
<tr>
<td>n=449,733</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014 NY</td>
<td>0.0019 (0.0018)</td>
<td>0.1986 (0.9400)</td>
<td>0.6927 (0.9494)</td>
<td>0.4395 (0.9907)</td>
</tr>
<tr>
<td>n=252,258</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014 WA</td>
<td>0.0002 (0.0030)</td>
<td>-11.0919 (199.5858)</td>
<td>15.2327 (259.6603)</td>
<td>18.5140 (315.9351)</td>
</tr>
<tr>
<td>n=119,769</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2015 PA</td>
<td>0.0116 (0.0045)</td>
<td>-0.3215 (0.3925)</td>
<td>0.1649 (0.1970)</td>
<td>-0.2050 (0.3364)</td>
</tr>
<tr>
<td>SUPREME COURT</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n=49,278</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2015 PHILLY MAYORAL</td>
<td>0.0078 (0.0023)</td>
<td>0.2097 (0.2676)</td>
<td>0.2504 (0.2159)</td>
<td>0.0239 (0.2739)</td>
</tr>
<tr>
<td>n=205,981</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2016 NC</td>
<td>0.0024 (0.0036)</td>
<td>NA</td>
<td>NA</td>
<td>-0.2869 (1.8542)</td>
</tr>
<tr>
<td>n=137,032</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2017 VA</td>
<td>0.0074 (0.0053)</td>
<td>NA</td>
<td>NA</td>
<td>0.6973 (0.6410)</td>
</tr>
<tr>
<td>n=67,971</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Washington was omitted from the main analysis since it is very likely that this analysis suffers from too weak an instrument. The first stage impact on upstream turnout was close to a perfect zero with a p-value 0.95.
4.11.2: Anonymized Canvasser Script and Training Documents

November 4, 2014

Governor: Mark Schauer;

US Senate: Gary Peters

[REDACTED] Persuasion Rap

Introduction

Hi, my name is ___ with [REDACTED]. Are you [name]? Great! We’re out today talking with folks in [insert community] about the election for Governor and the Michigan House.

QUESTIONS

Question 1

Are you planning to vote in the election?

[Do not record]

Question 2 (Voter ID Governor)

Thank you. If you were going to vote today in the election for Governor would you support for Republican Rick Snyder or Democrat Mark Schauer?

[Record Response: Snyder-R, Schauer-D, Unsure/Undecided]
MARK SCHAUER

PERSUASION AND ENDORSEMENTS

[REDACTED] is an independent organization that represents 120,000 Michiganders who want an economy that works for working people. We are not part of any political party or campaign and support candidates based on their record.

[IF SCHAUER]

We are also supporting Mark Schauer to be the next Governor. Thanks for your support!

*Hand over lit. Go to Peters Endorsement*

[IF UNDECIDED]

[REDACTED] has researched the issues and we have found that Mark Schauer has the strongest record of supporting working families and the issues important to the community. That is why he has the support of tens of thousands of working men and women in Michigan and will work to create good jobs and improve our schools.

So can we count on your vote for Mark Schauer for Governor?

*Do not record response. Hand over lit. Go to Peters Endorsement.*

[IF OTHER CANDIDATE]

I understand. How you vote is a personal decision. [REDACTED] has done the research on the issues and we believe that Mark Schauer has the strongest track record of getting things done for working people and will be the strongest leader for Michiganders.

*Hand over lit. Go to Peters Endorsement.*
US SENATE:

GARY PETERS

ENDORSEMENT

[REDACTED] has done the research and found that Gary Peters has the strongest record of fighting for Michigan. In Congress he helped deliver on the loans to save the auto industry jobs and start to turn around Michigan’s economy.

Go to Plan Making.

PLAN MAKING

[IF SCHAUER or UNDECIDED FOR STATE HOUSE]

Great! Thanks for your support. As you probably know, this election will determine the direction Michigan takes and turnout is going to be high.

The polls open at 7 am and close at 8 pm on Election Day. When do you normally go- on your way to work? When you get off? Just before dinner? [LISTEN FOR RESPONSE & FOLLOW UP IF APPROPRIATE. DRAW RESPONSE OUT OF VOTER.]

[FOR VOTERS WHO COMMIT TO TURNOUT]

Thanks. You know one of the best ways to increase Mark Schauer’s chances of winning is to talk to your family about why this election matters to you. Have you had a chance to talk to your family about this election? [LISTEN FOR RESPONSE & FOLLOW UP IF APPROPRIATE]

Sample follow up question: “How do they feel about the Governor’s race? Did you share your reasons for voting for Mark Schauer?”
Thanks, Goodnight.
PERSUASION TALKING POINTS

Crumbling Infrastructure: Fixing our roads and bridges is not Governor Snyder’s priority. The delay means an added $1.8 billion in costs for taxpayers, nearly $400 per year in unnecessary auto repairs for the average driver and countless hours of frustration. Experts say that one in four bridges rated as structurally deficient.

Mark Schauer will finally fix our roads and bridges by getting corporations to pay their fair share in taxes for the infrastructure they rely upon to conduct business.

Jobs and the Economy: Rick Snyder promised to focus on job creation, but in the last four years, Michigan has trailed the nation and neighboring states like Ohio in adding new jobs. Instead of creating jobs, Snyder has pursued divisive policies like “right to work” that hurt middle class Michiganders.

Mark Schauer has a record of defending Michigan jobs and workers. In Congress he supported the $80 billion investment to stabilize the auto industry. Since then, the money has been repaid and the Big Three are profitable. Schauer also wants to repeal the divisive Right to Work law, arguing that “Snyder cannot point to one job that’s been created as a result of right to work.”
Failing our Schools: In his first year in office, Governor Snyder cut over $1 billion in state school funding- more than $400 per pupil. As a result, class sizes have increased, learning environments have suffered and more than 50 school districts have experienced a credit down grade.

Mark Schauer will stop the diversion of Michigan School Aid fund dollars away from the classroom- nearly $400 per pupil that has resulted in increased class sizes and poorer learning environments, and threatens our finances.
Training Memo

Please distribute only to trainers

Opening the Rap: Canvassers have found it works better to use the words “upcoming election” with voters versus “out talking about issues”. They found that voters perceive “out talking about issues” as a longer conversation compared to “upcoming election”. Again work with folks on what approach works best for them – if there is push back on “upcoming election” it may be better to approach the voter with “out talking about issues in our community”

Inclusive Language: Reminding the voter that we are part of the same community and share the same interests facilitates a more collaborative conversation. For example, “That is why so many folks here in our community have decided that they are voting for Mark Schauer” assures the voter that we are looking at the issues from the same perspective. This approach assumes support. By contrast, exclusive language like “you” and “they” imply social distance and that you are attempting to convince the person to share the perspective.

Part of Something Bigger and Enthusiasm: It’s common for voters to want to feel like they are part of a larger effort to achieve a shared goal. Using language that conveys the point helps frame their decision to vote as part of a larger effort and makes them more likely to turnout. For example, “We have been talking to a lot of people who tell us that they are just tired of all of the negative attack ads that don’t do anything to fix the economy and other problems we are all facing. I am sure you probably feel the same way, right,” or “This is
going to be an important election, and we are expecting high turnout” are effective ways to convey that just like in school, all of the popular students are participating. It is for the same reason that it is important to show the voter that we are enthusiastic about supporting candidates who support our issues.

Clean IDs/Recording Responses: Capturing the initial response to issue and candidate ID questions is an important part of building the longer term political program. For example, know where we had the highest initial rate of support for candidate verses undecided voters informs where we prioritize future passes. While recording an ID after we have given some indication of our preferred candidate has value, the goal is to find those voters who are most likely to be with us.

Plan Making: An important part of plan making is that we walk voters through thinking about when and where they are going to go vote. Using time prompts and inclusive language, and effective plan making conversation helps the voter think about voting in the context of their own lives. For example, “It will be important to have plan on when you go to the polls. I like to go in the morning before work, but a lot of people like to go in the evening before dinner. What about you- when do you plan to head to the polls?” involves emphasizing everyday events that people plan their days around.

Talk To Your Family: A new part of the canvass is asking supportive voters to discuss the election with other voting members of their family. Research studies performed by the
[REDACTED] have shown that asking voters talk to family members about voting for Mark Schauer can increase overall vote probability for Schauer. The study found that getting supporters to talk to their family members about the election was over THREE TIMES MORE EFFECTIVE than existing practices of getting uncommitted household members to support a candidate.