5-8-2024

Essays on Censorship and Public Opinion in Authoritarian Regimes

Zirui Yang
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST.LOUIS

School of Arts & Science
Department of Political Science

Dissertation Examination Committee:
Margit Tavits, Chair
Deniz Aksoy
Taylor N. Carlson
Ted Enamorado
Haifeng Huang
Carly Wayne

Essays on Censorship and Public Opinion in Authoritarian Regimes

by
Zirui Yang

A dissertation presented to
Washington University in St. Louis
in partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

May 2024
St. Louis, Missouri

Table of Contents

List of Figures

List of Tables

Acknowledgments

This dissertation would not be possible without the unwavering support of my advisor, Margit Tavits. Throughout my graduate school, Margit has always been an incredible academic role model, pushing me to aim high and not settle for mediocrity. She is also a dedicated mentor, giving me detailed feedback on every single draft I would share with her. Beyond academia, Margit has been an ally in my personal life, helping me to overcome many obstacles I encountered as a minority scholar from a culturally conservative country. I could not have asked for a more committed, attentive, and compassionate advisor than Margit, and I am wholeheartedly grateful for everything she has done for me.

I am also deeply indebted to my dissertation committee members, Deniz Aksoy, Taylor Carlson, Ted Enamorado, Haifeng Huang, and Carly Wayne. Deniz has always supported me, helping me secure funding critical to my dissertation. Ted is the kindest person I have ever known in academia and has always encouraged me in the face of daunting challenges. Taylor and Carly provided some of the most insightful suggestions I ever received, helping me become a successful junior scholar, much like themselves. Haifeng generously taught me how to conduct research in China even before we knew each other well, providing me with opportunities I could not have otherwise received.

Academia is a difficult profession to navigate. Therefore, I am incredibly lucky and sincerely grateful to have three senior colleagues, Haohan Chen, Bryant Moy, and Luwei Ying, who gave me invaluable peer mentorship. The countless phone calls, coffees, and drinks with Haohan, Bryant, and Luwei guided me through all the secret codes in academia.

Graduate school, the first big challenge of academia, would have been so much more insufferable without the support and companionship of my wonderful cohort mates, particularly Afiq bin Oslan, Ben S. Noble, Ipek Ece Sener, and Jeremy Siow. Many thanks for their patience in brainstorming, commenting, and proofreading numerous drafts of this dissertation, as well as bearing my endless WhatsApp bombardment and occasional peer pressure against their anti-social personality. Additionally, I am grateful to other fellow members of our research group, Taylor Damann, Dahjin Kim, Rex Weiye Deng, and Patrick Edwards, who provided immense intellectual and emotional support throughout the entire period of my dissertation research.

Countless friends and colleagues inside and outside Washington University have given me invaluable support and suggestions for my dissertation and beyond. I thank (by alphabetical order) Joan Barceló, David Carter, Erin Carter, Keng-Chi Chang, Qianmiao Chen, Jane Esberg, King-wa Fu, Dimitar Gueorguiev, Rongbin Han, Mai Hassan, Ning He, Jae-Hee Jung, Jay Kao, Dixin Li, Jia Li, Xiaojun Li, Gechun Lin, Dongshu Liu, Peter Lorentzen, Fengming Lu, Jiaqi Lu, Xiaobo Lü, Christopher Lucas, Dan Mattingly, Jacob Montgomery, Lucia Motolina, Jen Pan, Miguel Pereira, Annamaria Prati, Juan Qian, Molly Roberts, Guillermo Rosas, Keith Schnakenberg, Li Shao, Shuyuan Shen, Matthew Shugart, Alex Siegel, Patrick Silva, Andy Stone, Cecilia Sui, Minh Trinh, Rory Truex, Josh Tucker, Erik Wang, Ye Wang, Anna Wilke, Jason Wu, Yiqing Xu, Eddie Yang, Hongshen Zhu, among many others.

I am also grateful for the generous financial support from several academic institutions that made my dissertation research possible. They are the Department of Political Science and the Weidenbaum Center on the Economy, Government, and Public Policy at Washington University in St. Louis, the Centennial Center for Political Science and Public Affairs at the American Political Science Association, the Chiang Ching-Kuo Foundation for International Scholarly Exchange, and the Institute for Humane Studies.

Moving to St. Louis, Missouri, from China is a formidable undertaking. Fortunately, I met a caring and supportive community that filled my time in St. Louis with cherished memories. I extend my heartfelt gratitude to Claire Chen, Paco Yip, Jiayang Chen, Weijia Cao, Alex Li, Mike Shen, Jianyong Jiang, and Vincent Lin for their friendship and companionship.

Finally, I am grateful for the love and support of my parents. I owe a special thank you to my father, Zezhu Yang, with whom I have shared many civil political discussions despite our differences in political views. It was one of these political debates that motivated this dissertation research. To my mother, Suqing Liang, I express my deepest gratitude for her understanding of my career path and life choices, which stems from her unconditional love for me.

This dissertation stands as a testament to the collective support and encouragement I have received from my colleagues, friends, and families. I am deeply thankful to each and every one of you for being an integral part of this remarkable journey.

Zirui Yang

Washington University in Saint Louis
May 2024

ABSTRACT OF THE DISSERTATION

Essays on Censorship and Public Opinion in Authoritarian Regimes

by

Zirui Yang

Doctor of Philosophy in Political Science

Washington University in St. Louis, 2024

Professor Margit Tavits, Chair

Conventional wisdom has long regarded censorship as a top-down, repressive tool for authoritarian governments to suppress political criticism and maintain regime stability, and therefore unpopular among the public. This dissertation proposes a novel normalization theory and presents various pieces of empirical evidence to challenge these conventional understandings of censorship and public opinion in authoritarian regimes. In the first chapter, I introduce the normalization theory and claim that the Chinese public does not necessarily perceive online censorship as repressive but as a normal part of Internet governance. Drawing from around 28 million censored posts on social media and two survey experiments, I demonstrate that, in addition to politically threatening content, non-political content is also censored on a substantial scale, which subsequently increases public support for censorship. In the second chapter, I further investigate why the public perceives censorship as normal by analyzing public participation in the censorship process. I propose a novel bottom-up perspective on censorship and demonstrate that censorship participation leads to higher public support for the censorship apparatus. In the third chapter, I challenge the conventional wisdom that authoritarian censorship tends to target positive exposure of foreign liberal democracies. Using a novel dataset of Chinese social media articles about liberal democracies from 2018 to 2022, I show that impeding mass

exposure to democratic institutions rather than defaming the West is the primary strategy of Chinese censorship. This study underscores the Chinese regime's lingering insecurity about public knowledge of liberal democratic systems. Taken together, this dissertation highlights how the normalization of censorship, ordinary citizens' participation, and strategic censorship of information about democracy contribute to maintaining public support and sustaining regime survival in autocracies.

Introduction

It was January 2023 when I first met Mr. Wen (pseudonym), a senior official working at the Cyberspace Administration (CA or *wangxinban*) in an affluent coastal province in China with thriving social media and tech companies. He was in charge of coordinating the censorship effort within the province. Merely a month and a half earlier, China witnessed unprecedented nationwide political protests against the draconian "Zero-COVID" policy, a public health emergency following the subsequent relaxing of COVID restrictions, and a sharp decrease in public trust in the government — a chaotic crisis rarely seen since the 1989 Tiananmen Square protests and massacre.

Mr. Wen has just attended a regular CA meeting on building "spiritual civilization" on the Internet (*wangluo jingshen wenming jianshe*), a euphemism for the Chinese Communist Party's (CCP) growing control over the cultural and entertainment sphere of the society. Despite mounting political pressure at the time and a spike in nationwide political censorship of the "White Paper Movement," a Soviet-style protest movement of mostly college students against government censorship, Mr. Wen seemed to have "bigger fish" to fry. In an almost tone-deaf manner, we chatted for three hours about seemingly politically benevolent things — how young entertainment stars are exerting "bad" cultural influence on TV shows, how inaccurate descriptions of Traditional Chinese Medicine have flooded the Internet, how the government should protect consumers from online advertisements,

and how well-intended life tips on social media can be misleading for senior users. Rarely did he mention how his department censored online criticism of the government and political protests during those months.

I knew he intentionally avoided politically sensitive topics, as misspeaking on political issues could seriously affect his promotion within the party rank. Just two days before, his less ambitious, and therefore more straightforward, colleague in the Public Security Department had already shown me their "army" of Internet censors dealing with "politically destabilizing" content. Yet, Mr. Wen was not lying. These seemingly trivial and harmless issues, such as censoring "misleading" entertainment stars and "inaccurate" information about Traditional Chinese Medicine, were indeed a substantive part of his daily work — our conversation was constantly interrupted by phone calls consulting him on these issues, and he showed me his future meeting schedules on "spiritual civilization," like the one he just attended. *Why, even during one of the most politically turbulent periods of the last three decades, does the Chinese censorship apparatus spend so much time and resources on seemingly harmless non-politically threatening content?*

Mr. Wen's absence from political censorship activities targeting online dissents and protests is not an isolated case. Many of his counterparts in other CAs across China were also not busy with political censorship either. Instead, they were responsible for running multiple propaganda campaigns to promote public participation in censorship by reporting online content (CNS 2020; Sina 2023). One communiqué states: "Cyberspace administrations across the country should place the promotion of online reporting in an important position in strengthening the comprehensive Internet management system" (Sina 2023). According to official statistics, the CAs in China have solicited over 172 million censorship requests in 2022 alone (Xinhua 2023). When the same resources could be directly spent on censoring

online content, *why does the Chinese censorship apparatus put so much effort into soliciting public participation in censorship?*

Even when censorship resources are directed toward politically sensitive topics, such as online discourse concerning foreign liberal democracies like the United States and Japan — China's primary international rival — the patterns of censorship remain puzzling. Throughout the COVID-19 pandemic, the efficacy of Chinese and American COVID vaccines become a major contestation both within China's domestic politics and on the international stage. Domestically, China stubbornly refused to import American vaccines even though most people believed them to be more effective, drawing criticisms at home and abroad. Internationally, China relentlessly promoted Chinese vaccines and carried out "vaccine diplomacy," claiming them to be as effective as American ones. However, despite the sensitivity of the issue, social media posts favoring American vaccines over Chinese ones were rarely censored. The Chinese government, the champion of vaccine nationalism, did not seem concerned about information on superior American vaccines spreading domestically. *What specific content about foreign democracies, then, does the Chinese censorship apparatus perceive as more threatening and subject to stricter censorship?*

These puzzling anecdotal observations in China, the world's most populous and influential authoritarian regime with the most sophisticated censorship apparatus in human history (Han 2018; King, Pan and Roberts 2013; Roberts 2018; Stockmann 2013), compel a reassessment of the conventional understanding of authoritarian censorship. Historically, censorship in authoritarian regimes has been construed as a top-down, repressive apparatus for autocrats to silence criticism and dissent (Gueorguiev and Malesky 2019; Miller 2018; Pop-Eleches and Way 2021), control the flow and narratives of unfavorable news (Rozenas and Stukal 2019; Shadmehr and Bernhardt 2015), and thwart collective action (King, Pan and Roberts 2013, 2014; Lorentzen 2014). It has been widely assumed

that government censorship primarily targets content critical of or destabilizing to the regime. As such, government censorship is viewed as an unpopular suppression of free speech, which leads to public backlash against the regime (Nabi 2014; Pan and Siegel 2020; Roberts 2018, 2020; Zhu and Fu 2021), and ordinary citizens often find creative ways to circumvent and resist government censorship activities (Chang et al. 2022; Gläßel and Paula 2020; Han 2018; Hobbs and Roberts 2018; Pan and Siegel 2020; Roberts 2018, 2020).

Yet, just like the anecdotes about Chinese censorship officials' puzzling behaviors, survey results and scholarly writings on public opinion and behavior toward censorship also reveal inconsistencies in the conventional wisdom of authoritarian censorship. First, public support for government censorship is much higher than the conventional wisdom suggests. Surveys across the world, and in particular in China, indicate a significant portion of the population within authoritarian regimes either displays apathy toward or actively supports governmental censorship efforts (Dickson 2016; Martin, Martins and Wood 2016; Nisbet, Kamenchuk and Dal 2017; Wang and Mark 2015; Wike and Simmons 2015). Second, the public's role extends beyond mere observation, as they frequently engage as active participants in the censorship process by flagging online content they deem objectionable (Cook 2019; Jiang 2021; Nimmo and Agranovich 2022; Tufekci 2017). This dissertation aims to explain these puzzling observations and update the scholarly understanding of authoritarian censorship.

Structure of the Dissertation

This dissertation consists of three independent studies, each constituting a chapter. In Chapter 1, I develop a novel normalization theory and argue that citizens support repressive apparatus like censorship because they do not view it as a form of political repression

but rather as a normal part of governance. I collected around 28 million censored posts on China's two largest social media platforms and conducted two original survey experiments to examine Chinese censorship practices and their effects on public perception of the censorship apparatus. I argue that citizens perceive censorship as normal because political content constitutes only a fraction of all censored content. Through systematic censorship of non-political content, citizens are less likely to associate censorship with political repression and are desensitized to future censorship events. Using the state-of-the-art deep learning models to classify the 28 million censored posts, I find that the majority are indeed unrelated to politically threatening topics, such as government criticism and collective expression. Moreover, both experiments show that respondents exposed to such non-political censorship exhibit significantly higher support for the regime and are more likely to believe that censorship is normal. These findings underscore the significance of normalization as a crucial channel for autocrats to maintain social control.

In Chapter 2, I further challenge the conventional wisdom that citizens perceive censorship as repressive. I explore a novel, bottom-up perspective of censorship in which ordinary citizens are encouraged to participate in censorship by reporting online content. Using an original survey and a creative experiment embedded in custom-engineered, simulated social media pages, I show that participation in censorship is prevalent among ordinary Chinese citizens, and such behavior, in turn, shapes their perception of and support for censorship. This study highlights the role of ordinary citizens in facilitating authoritarian control and further explains why repressive apparatus like censorship can be perceived as normal and popular with the population.

In Chapter 3, I argue that authoritarian regimes do not simply ban positive coverage of foreign liberal democracies, as a ban could compromise their credibility and lead to public backlash against their censorship. Instead, public discourse on democracies'

achievements is allowed, whereas conversations about democratic institutions, such as elections and legislative deliberations, are more stringently restricted. Using over 100,000 articles on Chinese social media and various natural language processing models, I find that whether the portrayal of democracies is positive or negative is a less significant predictor of censorship than whether it mentions democratic institutions. These findings demonstrate that the primary goal of censorship involves impeding mass exposure to democratic systems rather than merely defaming the West.

Together, this dissertation contributes to a deeper understanding of how modern authoritarian regimes, China in particular, manipulate information, operate repressive apparatus, and maintain popular support. It reconciles two central puzzles in the authoritarian politics literature: (1) how autocrats effectively control the public through repressive apparatus without provoking a backlash, and (2) the apparent support citizens show for overt repressive activities in authoritarian regimes. This dissertation proposes normalization as a novel and critical channel for authoritarian control. Such a channel is potentially difficult to undermine because citizens are desensitized and subconsciously accept the coercive apparatus such as censorship as normal.

The contribution of this dissertation also goes beyond authoritarian regimes. Internet regulation, social media censorship, and freedom of speech have become hotly debated issues in many democracies including the United States. In democracies, people worry about whether there is a slippery slope from censorship of offensive or socially harmful content to political speech. This dissertation might provide useful lessons on how citizens process online censorship and how to avoid unjustified political censorship in democracies.

Chapter 1

Normalization of Censorship: Evidence from China

1.1 Introduction

Government censorship in authoritarian regimes has traditionally been understood as a repressive tool to suppress political oppositions (Gueorguiev and Malesky 2019; Miller 2018; Pop-Eleches and Way 2021), filter unfavorable news (Rozenas and Stukal 2019; Shadmehr and Bernhardt 2015), and hinder collective action (King, Pan and Roberts 2013, 2014; Lorentzen 2014). As such, the public resists government censorship when they encounter it (Roberts 2018, 2020; Zhu and Fu 2021). For example, when citizens are exposed to censorship events, they express more anger and anti-regime sentiment (Pan and Siegel 2020; Roberts 2018, 137), discuss and search for more information on the censored topics (Nabi 2014; Pan and Siegel 2020; Roberts 2018, 143), show less support for Internet regulation and state media (Gläßel and Paula 2020; Roberts 2018, 144), and

even participate in protests against the regime (Boxell and Steinert-Threlkeld 2021). In light of these studies, Roberts (2020) posits that: "awareness of censorship [is] essential to resilience to censorship."

However, in China, where the scale of overt censorship activities is by far the largest around the world (Freedom House 2019; Gueorguiev and Malesky 2019; Han 2018; King, Pan and Roberts 2013; Miller 2018), surveys consistently find that Chinese citizens are either apathetic toward or supportive of the regime's censorship apparatus, even when they have experiences with censorship. For example, Dickson (2016, 71) reports that Chinese citizens who have experienced censorship are "rather blasé about it." Wang and Mark (2015) show that a majority (65.6%) of "censorship-aware respondents" are either neutral or supportive of Internet censorship. Studies on other authoritarian regimes such as Russia (Nisbet, Kamenchuk and Dal 2017) and Middle Eastern monarchies (Martin, Martins and Wood 2016; Wike and Simmons 2015) also find similar phenomena. *If awareness of censorship leads to backlash, why do many citizens in authoritarian regimes display such little resistance to the widespread use of overt censorship?*

This puzzle likely arises because the literature that developed the backlash argument primarily focuses on censorship of government criticism and collective action. Although politically threatening content has traditionally been understood as the prime target of censorship (Gueorguiev and Malesky 2019; King, Pan and Roberts 2013, 2014; Miller 2018; Roberts 2018), other seemingly harmless non-political content, such as popular culture (Nie 2021) and pornography (King, Pan and Roberts 2013), is also widely censored in authoritarian regimes. In this study, I argue that when the range of censored content extends beyond highly politically threatening content, such as collective action and government criticism, to other less politically threatening content, citizens are less likely to view censorship as political suppression. Rather, censorship becomes viewed as a normal government policy

that regulates both political speeches and apolitical content like entertainment, culture, and advertisement. I call such changes in perception *normalization of censorship*.

My argument about censorship normalization builds on the psychological theory of desensitization (Bartholow, Bushman and Sestir 2006; Carnagey, Anderson and Bushman 2007; Fanti et al. 2009). Conventionally, citizens react negatively to censorship of political content because it could signal that the government has something to hide and is not acting as a faithful agent (Lorentzen 2014; Roberts 2018; Shadmehr and Bernhardt 2015). When both politically threatening and less political messages are censored, it dilutes the probability that each censorship event contains valuable political information to discover government wrongdoings (Pan and Siegel 2020; Roberts 2020). In addition to dilution, censoring non-political content also increases citizens' exposure to censorship activities, which further facilitates the normalization process. As individuals are more frequently exposed to censorship activities, they are more likely to view censorship activities as normal events and not react as intensely. As a result, citizens' negative reactions toward censorship, such as anger and anti-regime sentiment, should be less likely to occur (Wang and Mark 2015).

To provide evidence for my theory of normalization, I first use observational data to illustrate that non-politically threatening content is being censored on a substantial scale. I collected and categorized over half a million censored articles from WeChat, China's largest social media platform, as well as over 27 million censored posts on Weibo, the second-largest platform in China. To ensure the accuracy of my findings, I employed both human coders and deep learning models, specifically the Chinese Bidirectional Encoder Representations from Transformers (BERT) with the Whole Word Masking model (Lu, Pan and Xu 2021). The results show that collective action, government criticism, and other government-related content only account for approximately 30% of all censored articles.

The majority of censored articles are non-politically threatening and include a wide range of non-political topics.

I then conducted two original survey experiments with similar designs in China, in 2020 and 2022 respectively, to test the effect of censoring non-politically threatening content on support for censorship and the regime. In both experiments, I randomly expose respondents to varying amounts of censorship of non-political content. Consistent with the normalization theory, respondents exposed to the censorship of both political and non-political content display significantly less backlash and greater support for the censorship apparatus and the regime, compared with those that are only exposed to the censorship of political content.

This study makes important contributions to the understanding of government censorship in authoritarian regimes. Although I am not the first to argue that authoritarian governments censor non-political content such as pornography (King, Pan and Roberts 2013) and popular culture (Esberg 2020), I challenge the centrality of political censorship in the literature. The results of this study highlight that the censorship of non-political content is crucial to understanding how political censorship can work effectively. My results also bridge the gap between the seemingly contradictory observations that on one hand, censorship awareness will lead to backlash, and on the other hand, citizens are numb to the massive overt censorship activities in China. Finally, I expand the existing understanding that overt censorship is effective primarily because it creates fear and deters dissent (Roberts 2018, 2020), whereas covert forms of censorship, such as "friction" and "flooding," are more effective in avoiding public backlash (Miller 2018; Roberts 2018, 2020; Stukal, Sanovich, Bonneau and Tucker 2022). Overt censorship, when applied broadly, might desensitize the public to censorship which subsequently reduces public backlash.

A broader implication of these findings concerns the dynamic of authoritarian control. A wealth of scholarship has investigated how authoritarian governments persuade and threaten people through propaganda, silence dissent through censorship, and prevent uprisings through repression (Arendt 1976; Cantoni et al. 2017; Chen and Xu 2017; Dickson 2016; Guriev and Treisman 2015; Huang 2018; Shadmehr and Bernhardt 2015; Svolik 2012; Young 2019). My study suggests that there might be an additional channel through which authoritarian regimes achieve social control: the normalization of coercive policies. Such a channel is potentially difficult to undermine because citizens are desensitized and subconsciously accept the coercive policy as normal.

1.2 Normalization of Censorship: A Theory

The central argument of this study is that, when the range of censorship expands beyond highly politically threatening content, such as collective action and government criticism, to other less politically threatening content, it normalizes the censorship apparatus and desensitizes citizens to censorship activities. As a result, backlash against both the censorship apparatus and the regime is less likely to happen. Such normalizing effects are primarily achieved by diluting the proportion of politically threatening content among censorship targets. In addition to dilution, censoring non-political content also increases citizens' exposure to censorship activities, which further facilitates the normalization process.

1.2.1 Diluting the Proportion of Politically Threatening Content

Backlash to censorship happens when citizens are aware of censorship activities and care about what has been censored (Chen and Yang 2019; Roberts 2020). The extent to which citizens care about censorship activities depends on the nature of the censored

content. Conventionally, the government tends to target the most politically threatening information, such as messages with collective action potential and government criticism (Lorentzen 2014; King, Pan and Roberts 2013, 2014; Shadmehr and Bernhardt 2015). Hence, censorship could be seen as a signal that the regime has something to hide and is not acting as a faithful agent for the citizens (Lorentzen 2014; Roberts 2018; Shadmehr and Bernhardt 2015). It indicates abnormality and potential government wrongdoings. As a result, citizens will pay even closer attention to the censored information to find out what has been hidden from them. Such an effect is called the Streisand effect: the act of censorship drawing even more attention to the event that the government initially tried to cover up (Roberts 2020). Given the difficulty of completely covering up information on the Internet (Roberts 2018), once the citizens uncover the censored information, the anger toward the government will be magnified. Consistent with this logic, several recent studies have found evidence of backlash against censorship of political content (Pan and Siegel 2020; Roberts 2018, 2020).

For such backlash to occur, however, it is critical that citizens believe censorship is abnormal and that censored information is valuable for discovering government wrongdoings. If citizens view censorship as normal, they are less likely to pay attention to censorship events in the first place. Therefore, a Streisand effect is less likely to occur.

According to psychological research on desensitization, when subjects' categorization and expectation of a stimulus are shifted from negative to neutral (or even positive), they will be less sensitive to the stimulus and their negative reactions to the stimulus will be diminished (Carnagey, Anderson and Bushman 2007; Efran and Marcia 1967; Goldfried 1971; Marcia, Rubin and Efran 1969). For example, playing violent video games or watching violent movies will expose subjects to initially negative stimuli (i.e. violence) in a positive emotional context. As a result, subjects will change their normative evaluation of violence

and decrease their attention to violent events (Carnagey, Anderson and Bushman 2007). Similarly, if citizens are exposed to censorship in a neutral or positive context, they will view censorship as normal and be less likely to pay attention to censorship events. As a consequence, subsequent backlash against censorship is less likely to happen.

Under what conditions will citizens be exposed to censorship in a neutral or even positive context? Direct interaction with censorship, such as having one's own message censored, can evoke strong reactions, especially when they find an online community of outspoken users who are also angry about the censorship event (Pan and Siegel 2020; Zhu and Fu 2021). In contrast, indirect interaction with censorship, such as observing a web page blocked or discussing censorship online, may yield less negative responses. When users come across deleted posts, they rely on available information, such as the post's title, author, and other users' comments and reactions, to form their beliefs about the censorship event.[1]

If censored content consistently concerns politically threatening topics, users tend to associate censorship events with negative news and government wrongdoings. In contrast, censorship of non-political content dilutes this negative image. When users encounter censorship of non-political content, they update their beliefs and are less likely to associate censorship with hiding politically valuable information or silencing political dissent. Consequently, they are also less likely to witness an angry public reaction against censorship.

Moreover, users frequently engage in discussions about censorship in online conversations, often in neutral and apolitical contexts (Han 2018). For instance, in February 2020, fans of different entertainment stars engaged in heated debates that often involved advocating for

[1]The extent of remaining information post-censorship varies among individual websites and social media platforms. For instance, on WeChat, users can still view the title of a deleted article, but they only discover the content is censored if they attempt to access the full article. Sina Weibo employs a more diverse range of censorship techniques Miller (2018).

the censorship of opposing posts. Such conversations imply that censorship can extend to non-political content related to entertainment stars. Observers of these discussions may be less inclined to view censorship negatively and connect it with political repression.

According to existing surveys, only 9% of citizens have directly experienced censorship (Dickson 2016), even though 69.5% of citizens are aware of censorship (Wang and Mark 2015). Hence, direct interaction with censorship is relatively rare, and most citizens form their beliefs about censorship through indirect interactions. Thus, the range of censored content significantly influences citizens' attitudes toward censorship. When seemingly harmless non-political topics are included among censorship targets, the effectiveness of censorship as a signal for government wrongdoing diminishes. As a result, citizens adjust their beliefs about censored information and are less likely to exhibit significant backlash when encountering censorship. In short, including non-political content among censorship targets dilutes the proportion of politically threatening content, and changes citizens' belief that censorship is abnormal and hides politically valuable information such as government wrongdoings. Consequently, negative reactions to censorship of political content are less likely to occur.

1.2.2 Increasing Citizens' Exposure to Censorship

In addition to dilution, censoring non-political content increases the frequency of citizens' exposure to censorship, which further facilitates the normalization process. As stated above, the belief that censorship is abnormal is critical for backlash to occur. If the chance of encountering censorship increases in citizens' daily lives, it is more likely for them to view censorship as normal and not pay too much attention to it.

A deeper look into the psychological mechanism suggests that such a desensitizing effect is due to the blunted reactions after repeated exposure to similar stimuli. Initially, a negative stimulus, such as violence or repression, arouses cognitive, physiological, and emotional responses (Anderson et al. 2010; Bartholow, Bushman and Sestir 2006; Carnagey, Anderson and Bushman 2007). Repeated exposure to the same stimulus, even over a short period of time, leads to blunted evaluative categorization and elimination of physiological and emotional reactions (Bartholow, Bushman and Sestir 2006; Carnagey, Anderson and Bushman 2007; Fanti et al. 2009). Similarly, although the initial exposure to censorship might arouse intense cognitive and emotional reactions, such as anger and resentment, such cognitive and emotional responses should be less likely to occur as individuals are more frequently exposed to censorship activities. As a result, citizens are more likely to regard censorship as normal.

The normalizing effect of increased exposures further facilitates dilution. Because citizens regard censorship as normal and common in daily life, they would not deliberately avoid it in conversation as they do with other sensitive topics. The example above about fans of entertainment stars illustrates how censorship could be a normal topic that is frequently mentioned in online conversations. The effect of increased exposure becomes a positive circle that reinforces the belief that censorship is normal.

An important implication of the normalization theory is that normalization is not merely a combination of popular and unpopular events, but rather a systematic shift in public perception of censorship, which affects their downstream beliefs of censored content and subsequent behaviors when encountering censorship activities. Once individuals believe censorship is normal, they are less likely to pay attention to censored content and are more likely to dismiss most censored content as not valuable or "deserved to be censored." As

such, individuals who believe censorship is normal should show higher support for the censorship apparatus as a whole, not just the non-political part.

1.2.3 Empirical Expectations

To summarize the empirical expectations of the theoretical arguments laid out above, I hypothesize that *citizens exposed to censorship of both political and non-political content, will display greater support for both the censorship apparatus (H1) and the regime (H2), compared with citizens exposed to censorship of political content only.*

Before I test the two main hypotheses, however, I need to demonstrate the possibility that censorship normalization happens in China. In the next section, I combine the existing literature on the Chinese censorship regime with my fieldwork conducted from November 2022 to February 2023 in China to illustrate the institutional environment for the normalization of censorship. I then use data on censored articles to explore one observable implication of the normalization theory: non-political content accounts for a large proportion of all censored content, i.e., Pr(Non-Political Content|All Censored Content) is high. To be clear, this is not to say that non-political content is more likely to be censored than political content, i.e., Pr(Censorship|Non-Political Content) $>$ Pr(Censorship|Political Content), or vice versa. I do not test which category is more likely to be censored. Instead, I aim to demonstrate that censorship of non-political content happens on a substantial scale.

1.3 Institutional Development of China's Censorship

Although online censorship occurs in almost every authoritarian regime, the range and scale of government censorship in China is by far the largest (Freedom House 2019). To

achieve normalization explained in the theory section, the regime often needs a relatively high capacity censorship apparatus that not only serves the function of political repression but also manages other non-politically threatening content.

Before the 2010s, censorship power in China was fragmented among different authorities (Alshabah 2016; Han 2018; Roberts 2018).[2] During this period, while the Chinese state has been relatively successful in curbing political threats from the Internet to regime stability similar to the Arab Spring, the regulatory fragmentation has led to the inability to exert extensive control over the Internet (Yang 2009). This is reflected in earlier studies showing that Chinese censorship primarily targeted the most dangerous content with collective action potentials while allowing most other online expressions (King, Pan and Roberts 2013, 2014).

In 2014, the Chinese leader Xi Jinping launched an effort to unify and centralize the administrative power of cyberspace, establishing the Central Leading Group for Cybersecurity and Informatization, which is directly chaired by Xi Jinping himself (Tai and Fu 2020). The Leading Group gained control over the Cyberspace Administration of China (CAC), which used to be primarily led by the State Council, and rapidly expanded its regulatory responsibility, gabbing power from other ministries under the State Council.[3] The centralization of censorship power had resulted in a more aggressive approach to Internet censorship in China. In addition to banning more political content than previously understood (Gueorguiev and Malesky 2019; Miller 2018), many seemingly harmless posts such as tabloid gossip were also censored (Cairns and Carlson 2016; Han 2018; Huang 2017; Ng 2015).

[2]This period is often referred to as the period where "nine dragons reign together." Censorship power was separated among departments ranging from propaganda, public security, military, and national security, to education, industry, and labor.

[3]According to my conversations with censorship officials during my fieldwork, the then director of the CAC, Lu Wei, played a critical role in the power-grabbing process of the CAC. His strongmen style leadership helped the CAC to become the central coordinator of the entire censorship apparatus.

Once-tolerated platforms focused on apolitical topics, including entertainment and dating applications, faced new restrictions (Freedom House 2019). Information and discussions on subjects like the economy that had traditionally been given freer rein became more systematically censored (Tai and Fu 2020).

Meanwhile, the government's narrative had also become more explicit about its intention to expand the censorship apparatus beyond safeguarding political stability. Before the 2010s, defending national security against foreign subversion had been a central theme of government documents related to Internet censorship, with only a few exceptions such as campaigns against online pornography (Han 2018). This is consistent with the traditional view of censorship as a tool of political repression and a showcase of regime strength. As the censorship power centralized and expanded during the early 2010s, official propaganda rhetoric became more diverse and less politically threatening. Instead, it tended to emphasize the less politically sensitive side of the Internet and highlight the need for extensive Internet regulations.

In the late 2010s and early 2020s, as political and technological competition with the United States and other Western powers intensified, the CAC doubled down on the national security rhetoric, citing foreign subversion threats, while also emphasizing the non-political benefits the authority provided such as digitization and Internet civility.[4] Combining the different pieces of anecdotes above, although censorship activities have been expanding rapidly in recent years, the image of censorship that the Chinese government tries to present has become more benevolent and less politically repressive. The narratives of the central government suggest an association between the expansion and the normalization of censorship.

[4]Such parallel messaging is reflected in both official documents published by the CAC and my conversations with central and local censorship officials during my fieldwork.

However, it is worth acknowledging that, even after the centralization of censorship power, the Chinese censorship apparatus is still far from a monolithic system in which every censorship activity stems from the command of a central authority. Instead, it is a comprehensive apparatus that involves many significant decision makers such as local governments (Lorentzen 2014) and social media companies (Han 2018; Lv and Luo 2018; Miller 2018), which also exert certain levels of influence over what content is censored. The purpose of this study is not to show that the normalization of censorship is a grand strategy designed and implemented by the central government alone. Rather, I aim to highlight the recent expansion of censorship power and increase in censorship activities in China, and its downstream normalizing effects on public perception of the censorship apparatus.

1.4 The Nature of Censored Content: Text Analysis

To better illustrate the possibility that censorship normalization happens in China, I use text analysis to more rigorously show that the censorship of less politically threatening content occurs on a substantial scale. I collect censored articles from WeChat and Sina Weibo, the two largest social media platforms in China. I then classify the censored articles into nine different topic categories, including three highly political, two moderately political, and four non-political categories. The main outcome of interest is the proportion of censored articles by topic category. To ensure the reliability of the categorization process, I use both human coders and natural language processing models, specifically the Chinese Bidirectional Encoder Representations from Transformers (BERT) with the Whole Word Masking model (Lu, Pan and Xu 2021).

1.4.1 Data Source

My observational study relies on three distinct sources of censorship data from China. Two of these sources are the WeChatScope and WeiboScope datasets, which were collected by a research team at the University of Hong Kong (Tai and Fu 2020; Zhu and Fu 2021). These datasets monitor over 4,000 WeChat public accounts and more than 118,000 Sina Weibo users, respectively. The third dataset, FreeWeChat, was gathered from the non-governmental organization GreatFire.org, which monitors over 34,000 WeChat public accounts in real time. WeChat public accounts are similar to Facebook public pages or Telegram public groups, while Sina Weibo is similar to Twitter. Given that WeChat public accounts have a large number of subscribers and posts on Weibo are easily disseminated, both platforms are prime targets of Chinese censorship. Table 1.1 summarizes the three data sources.

Table 1.1: Date Sources for Observational Analysis

	WeChatScope	FreeWeChat	WeiboScope
Platform	WeChat	WeChat	Sina Weibo
Content Source	Public Accounts	Public Accounts	Users
Number of Censored Posts/Articles	15,872	533,707	27 million
Data Start Date	2018-03-01	2016-04-25	2021-05-01
Data End Date	2020-05-09	2022-12-26	2022-06-30
Data Source	HKU	GreatFire.org	HKU

Moreover, both WeChat and Sina Weibo are ideal platforms for analyzing the implication of government censorship on citizens' attitudes because they are the two most popular social media in China. As such, censorship on WeChat and Weibo influences a large proportion of the Chinese population, and most Chinese are likely to form their beliefs about censorship via WeChat and Weibo. Using the three data sources provides a hard case to illustrate

the existence of large-scale censorship of non-political content. Because all three projects are designed to capture government censorship as conventionally understood, political accounts are over-represented in the sample (Tai and Fu 2020; Zhu and Fu 2021).[5] Therefore, even if the selected WeChat public accounts and Weibo users might not be representative, the bias is likely to be in the opposite direction of my theoretical expectations.

One caveat to all three data sources is that, like most quantitative censorship data, it only includes *post hoc* censorship. The articles or posts need to be published on WeChat or Weibo before they can be manually censored and recorded in the database. However, an article has to pass *ex ante* censorship barriers such as "keyword blocking" and "The Great Firewall of China" before it can be posted online (King, Pan and Roberts 2013). According to previous analyses of taboo keywords, *ex ante* censorship is focused overwhelmingly on political topics (Han 2018; Ng 2015). Thus, political content might have a higher bar for publication than non-political content. Nonetheless, citizens have creative ways to bypass keyword blocking and *post hoc* censorship is the most extensive form of censorship (Han 2018; King, Pan and Roberts 2013). As such, I focus on *post hoc* censorship in this analysis.

1.4.2 Categorization of Censored Articles

The primary objective of the categorization is to determine the proportion of highly political, moderately political, and non-political content. To define highly political content, I refer to the existing literature on censorship, which identifies three main highly political categories: (1) collective actions, including protests, strikes, rightful resistance, and other collective civil disobedience (King, Pan and Roberts 2013, 2014); (2) government criticism,

[5]The WeChatScope and WeiboScope projects primarily include accounts related to social and political news or commentary. It also samples influential public accounts including (a) public accounts for the government and the Communist Party; (b) high-ranked accounts; and (c) accounts with article links posted on a major discussion forum or indexed by the Baidu search engine (Tai and Fu 2020; Zhu and Fu 2021).

which entails criticisms of the central or local government, as well as state-owned enterprises (Gueorguiev and Malesky 2019; Lu, Pan and Xu 2021; Tai and Fu 2020); and (3) other government-related content, such as discussion of political leaders and political rumors (Miller 2018), or pro-government content that nonetheless mentions sensitive political topics or figures (King, Pan and Roberts 2013, 2014, 2017; Lu, Pan and Xu 2021; Stukal et al. 2022).

In addition to the highly political categories, there are two moderately political categories that may still be sensitive to the government: business and foreign events. The business category includes articles discussing private companies, while the foreign events category covers content related to foreign countries, provided that it does not directly reference China or the Chinese government's economic and foreign policies. Examples of business content include investment tips in the stock market, and examples of foreign events include discussions of the domestic politics of foreign countries. Notably, none of the existing censorship studies consider these categories as politically threatening to the Chinese government (Gueorguiev and Malesky 2019; King, Pan and Roberts 2013; Miller 2018; Tai and Fu 2020). However, some studies on authoritarian propaganda imply that business and foreign news might be relevant for autocrats' legitimacy and political survival (Mattingly and Yao 2022; Rozenas and Stukal 2019). As such, I first follow the censorship literature and only treat collective action and government-related content as political, and then include the two additional categories to provide a more conservative measure of non-political content.

Lastly, I track four non-political categories that are unlikely to pose a threat to the Chinese government: entertainment, advertisement, culture, and other content. Examples of these non-political categories include tabloid gossip, product promotions, and stories of cultural figures. To ensure consistency and accuracy, I adopt the coding rubrics primarily from

Miller (2018), which provides a comprehensive and up-to-date categorization of censored content in China. Notably, I make one key modification: the nine categories are mutually exclusive. This simplifies both the categorization process and the interpretation of results.[6]

To categorize the vast number of censored articles, I first employ human coders to annotate a training set. I randomly sampled 2,500 articles from WeChatScope, and 5,000 articles from the remaining two data sources, stratified by the creation year. Two native Chinese coders have coded the training set independently. The Cohen's κ between the two coders is 0.80, higher than the commonly applied criteria of 0.70 for inter-coder reliability tests. In cases where the two coders disagreed, the author acted as an arbitrator to settle the dispute.

To select the best classification model, I utilized the training data to evaluate the performance of nine different machine-learning models through out-of-sample cross-validation.[7] The fine-tuned pre-trained Chinese BERT model is by far the best-performing model, achieving an in-sample accuracy of 0.96 and an out-sample balanced macro F1 score of 0.72. Finally, I use the fine-tuned Chinese BERT model to predict the categories of censored content from all three sources. For more details about the performance of BERT and alternative models, please refer to Appendix E.

1.4.3 Results

Table A.15 reports the predicted proportion of censored articles and posts classified by topic category across the three different data sources. Consistent with the empirical expectations,

[6]For the detailed explanation of each topic category and coding process, see Appendix D.

[7]These models include a fine-tuned pre-trained Chinese BERT with the Whole Word Masking model, a logistic regression model with a ridge estimator, a pattern learning and matching (PaLM) model, an ensemble classifier model, a random forest model, a decision tree model, an extreme gradient boosting (XGBoost) model, a neural network model, and a word embedding model using Word2vec.

a substantial proportion of censored articles are unrelated to politically threatening topics. Let me first focus on the highly political categories highlighted by the existing literature. As illustrated in the top section of Table A.15, only about 30% of all censored articles across all three data sources pertain to collective action and government-related topics. In contrast, the remaining content that is less politically threatening constitutes the vast majority of censored articles on both WeChat and Sina Weibo. It is important to note that censorship of collective action and government criticism, which are the two categories that have the highest likelihood of generating popular backlash against the regime (Pan and Siegel 2020; Roberts 2020), only occurs about 20% of the time when citizens experience censorship events. In other words, it is much more common for Chinese citizens to encounter censorship of non-political content than censorship of government criticism and collective action. This would significantly lower their expectations that censorship events primarily target political oppositions and are useful for uncovering government wrongdoings. As a consequence, citizens would be more likely to believe that censorship is normal.

Even when considering a broader definition of political content that includes moderately political categories such as business and foreign events, which, as previously explained, explicitly exclude content related to the Chinese government, political content still constitutes only approximately half of all censored content. On Weibo, censored political content is particularly scarce, with only one-third of the censored content in WeiboScope being categorized as highly or moderately political. Additionally, the censorship of non-political content is not concentrated on a single category. For instance, the WeChatScope data recorded around 16.87% of entertainment content, 10.69% of advertisements, and 12.90% of cultural, local, and religious content. The other two censorship datasets also present similar patterns. These findings suggest that the censorship of non-political content is

Table 1.2: Predicted Proportion of Censored Articles/Posts by Topic Category

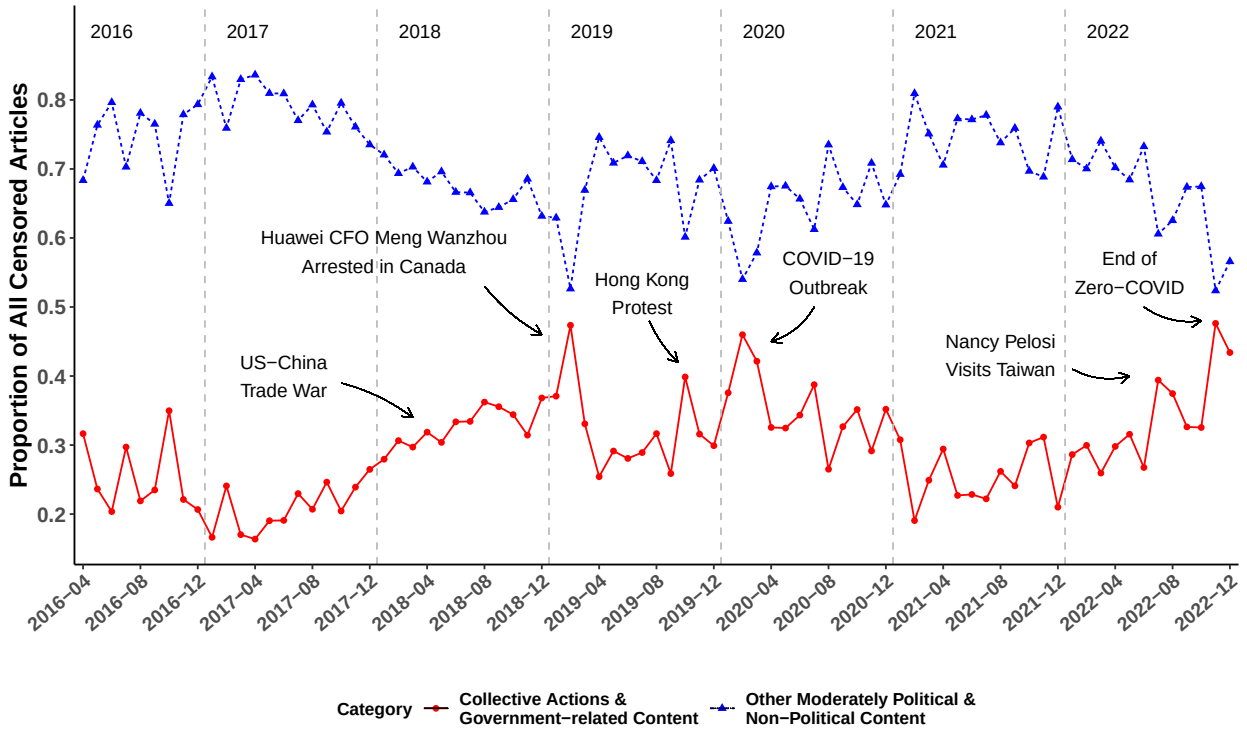
General Category	Specific Category	WeChatScope	FreeWeChat	WeiboScope
Highly Political	Collective Action	1.78%	0.78%	0.52%
	Govt Criticism	18.23%	11.79%	11.06%
	Other Govt-related	11.84%	14.03%	13.41%
	Total	31.85%	26.60%	24.99%
Moderately Political	Business	13.87%	9.95%	6.87%
	Foreign	5.89%	6.96%	5.12%
	Total	19.76%	16.91%	11.99%
Non-Political	Entertainment	16.87%	19.34%	21.71%
	Advertisement	10.69%	10.90%	16.32%
	Culture	12.90%	20.63%	17.72%
	Others	7.92%	5.61%	7.28%
	Total	48.38%	56.48%	63.03%

Notes: WeChatScope data contains 15,872 censored articles. FreeWeChat data contains 533,707 censored articles. WeiboScope data contains approximately 27 million censored posts.

comprehensive and broad, rather than being limited to a specific type of non-political content, such as popular culture (Esberg 2020) or pornography (King, Pan and Roberts 2013), indicating that its objectives extend beyond merely content moderating inappropriate content to a more systematic control over less politically sensitive areas.[8]

The observation that a considerable percentage of censored articles are less politically threatening in nature is consistently observed over time. Figure 1.1 presents the time series data depicting the proportion of collective action, government-related articles, and other less politically threatening topic categories. Even during the initial stages of the COVID-19 outbreak (February 2020) and later, during COVID policy U-turns (November 2022), when government suppression of online criticism was at its peak, non-politically threatening content still constituted more than half of all censored articles.

[8]Appendix D.3 uses structural topic models to provide further details regarding the content in each specific topic category.



Notes: All three data sources are weighted equally. Other moderately political and non-political content includes Business, Foreign Events, Entertainment, Advertisement, Cultures, and Others.

Figure 1.1: Time Series of All Censored Articles/Posts by Topic Category

To validate the text analysis results, I randomly sampled 1,000 machine-categorized non-political articles and hired a different undergraduate research assistant from China to identify whether there were any hidden or explicit political messages in them. The results of this validation exercise suggest that 91.2% of the machine-categorized non-political content is indeed irrelevant to politically threatening topics. Even after correcting for the 10% miscategorization, a majority of the censored content will still be non-politically threatening. This also alleviates the concern of political euphemism or other strategies citizens use to circumvent censorship, which could potentially mislead the text analysis results (King, Pan and Roberts 2013; Han 2018).

The implications of the empirical results are two-fold. First, the results from both WeChat and Weibo censorship echo the narrative in the previous section showing that the Chinese government has spent considerable effort in censoring non-politically threatening content while downplaying suppression of political discussion. The fact that collective action and government-related articles account for a minority of all censored content strengthens the government's claim that censorship is normal and benevolent. Besides, the fact that there is not a clear pattern for non-political content implies that the censorship of non-political content is broad and comprehensive rather than narrow and focused on one specific type of non-political content such as popular culture (Esberg 2020) or pornography (King, Pan and Roberts 2013). This keeps the red line fuzzy and gives the government leeway to include politically threatening content in broadly targeted censorship activities without drawing too much attention to the political content (Han 2018).

Second, the results imply that it is common for Chinese citizens to encounter censorship of non-political content. As a result, their expectations of censorship outcomes are likely to be shaped by these experiences of censorship of non-political content, which provides favorable conditions for the desensitization process to take place.

1.5 Normalizing Effects of Censoring Non-Political Content: Survey Experiments

The previous section illustrates the possibility that censorship normalization happens by showing the existence of large-scale censorship of less politically threatening content. To test the two main hypotheses, I conducted two original survey experiments in China with similar experimental designs. In both experiments, I manipulated the topics of censorship targets and the frequency of censorship activities that participants were exposed to. I

then use their *ex post* evaluations of the censorship apparatus and the Chinese regime to measure the normalizing effects of censorship.

1.5.1 Participants

The first survey experiment was conducted in December 2020 with 612 respondents and the second was conducted in December 2022 with 3,314 respondents.[9] The participants in both experiments were recruited from a Chinese survey platform and then directed to an American-based website, Qualtrics, where they completed the survey anonymously. Previous research has shown that using online platforms is a reliable way to recruit participants for survey experiments (Mullinix et al. 2015) and Chinese online platforms might even be better because political opinion surveys are relatively rare in China and participants are less likely to be professional political survey takers (Huang 2018; Huang and Yeh 2019).[10] To further ensure sample quality, I used attention checks to screen the respondents at the beginning of the surveys. Both samples cover a wide range of socioeconomic backgrounds and are similar to the Chinese demographic in terms of gender, age, and regional distributions. However, like many other online surveys in China, the samples are richer and better educated.[11]

1.5.2 Experimental Design

The purpose of the experiments is to compare the downstream effects of censorship that primarily targets politically threatening content with censorship that targets *both political*

[9]Both experiments were approved by the Institutional Review Board (IRB ID: 202011148) at the researcher's home institution and were pre-registered on the Open Science Framework.

[10]At the end of the survey, respondents were allowed to leave comments on the survey questions. The comments suggest that participants were excited rather than scared by political surveys.

[11]See Appendix A for more descriptive statistics.

and non-political content. Both experiments have similar designs and include three components. First, participants answered pre-treatment questions about their socioeconomic backgrounds and political predispositions. Second, participants were randomly assigned to either a control group, where they were exposed to only censorship of political content, or a treatment group, where they were exposed to censorship of both political and non-political content.[12] Finally, respondents answered post-treatment questions about their attitudes toward the censorship apparatus and the regime.

To expose participants to censorship, I asked respondents to read ten snippets of WeChat articles, presented one at a time with only the title and the first few lines.[13] Some of the snippets were labeled as censored by WeChat. These censorship labels primed the respondents that certain articles were censored, and no further information was provided to the respondents. Figure 1.2 shows one of the snippets with and without a censorship label.[14] Among the ten snippets, six were about non-political topics, whereas four were about government criticism and collective action. The order of the snippets was randomized.

I validated the appropriateness of my choice of article snippets in two ways. First, I consulted a panel of China scholars. None of them thought that any of the snippets were absurd or unreasonable. Second, I also asked the respondents about their interest in reading the full article after they read each snippet. If a snippet is particularly inappropriate, then this would be reflected in an unusually high or low level of interest. None of the snippets

[12] In the second experiment, there was an additional blank control group where respondents were not shown any censorship at all. For consistency between the two experiments, I only focus on the treatment and control groups in the main analyses. See section 5.5 for more discussion of the blank control group.

[13] The snippets are screenshots of real articles censored by WeChat. They only include the first couple of lines and do not reveal the full content of the articles. For the first experiment, I selected snippets from the WeChatScope dataset. For the second experiment, I selected snippets from the FreeWeChat dataset. The selection process was systematic. The details of the article snippets can be found in Appendix C.

[14] In reality, if an article is censored by WeChat, users can only see the title of that article but not the first couple of lines. In this experiment, to ensure the only difference between the treatment and control groups is the censorship label, I use the same snippets for both groups.

Notes: The censorship label reads: This article was blocked by WeChat due to violations of Internet law.

Figure 1.2: Example Snippet with and without Censorship Label in the Experiment

stands out as exceptionally high or low in terms of the level of interest. Hence, the experimental results are not driven by "inappropriate" or unreasonable choice of articles.

In the control group, three out of the four political snippets were labeled as censored by WeChat, whereas none of the non-political snippets were labeled. This primed the respondents in the control group that censorship primarily targets politically threatening content. In the treatment group, three out of the six non-political snippets were also labeled as censored by WeChat in addition to the three political snippets in the control group. In other words, there were a total of six snippets that were labeled in the treatment group and three in the control. This primed the respondents in the treatment group that both political and non-political content could be censored. This also means that respondents in the treatment group are exposed to twice as many censored snippets as those in the control group. Hence, the treatment reflects both dilution and increased exposure laid

out in the theory. Table 1.3 summarizes when and where the censorship label occurred in the treatment and control groups. Labeled snippets remained constant across subjects in respective groups, though as mentioned, their orders were randomized (i.e., labeled snippets might occur at any position).

Table 1.3: Explanation of Treatment

Snippet #	Topic Category	Control Group	Treatment Group
1	Political	Censorship Label	Censorship Label
2	Political	Censorship Label	Censorship Label
3	Political	Censorship Label	Censorship Label
4	Political		
5	Non-Political		Censorship Label
6	Non-Political		Censorship Label
7	Non-Political		Censorship Label
8	Non-Political		
9	Non-Political		
10	Non-Political		

Notes: The order of the snippets was randomized. The treatment group was exposed to additional censorship of non-political snippets.

1.5.3 Measurement

In both survey experiments, to measure *support for censorship apparatus*, I asked the respondents whether they agree or disagree that "the government should actively control the Internet and remove content that it deems inappropriate." This is a direct test of hypothesis 1 and the wording is an adaptation inspired by similar questions in the limited number of existing studies on public attitude toward censorship in China (Dickson 2016; Roberts 2018; Wang and Mark 2015). I avoided directly using the word "censorship" because it might be offputting to respondents. To measure regime support, I directly borrowed questions measuring assessment of the government, *overall satisfaction*, and *willingness to protest* from

Huang (2018). For the assessment of the government, I asked separately about the *central government* and *local government*. Because discussing local government is less sensitive than the central government, it alleviates potential social desirability bias and ceiling effects problems. For the last question, I expect respondents in the treatment group to be less willing to participate in protests. All outcome variables were measured on a five-point scale.

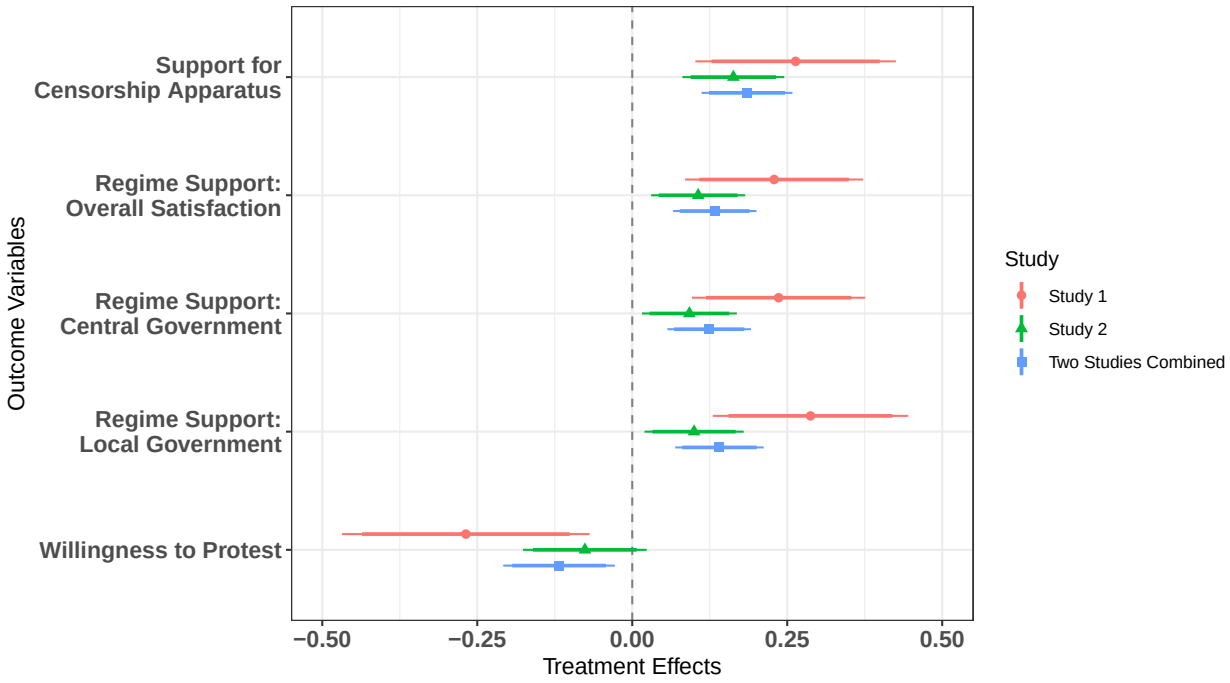
I used eight pre-treatment covariates to check the balance between the treatment and control groups. They are also included in the regression analyses. Among the covariates, four are demographic variables, including *age group*, *income*, *gender*, and *education*. The remaining four covariates are *ideology*, *party membership*, *political interests*, and *social media usage*, which measure participants' political predispositions and Internet usage.

1.5.4 Results

Figure 1.3 reports the difference in means between the treatment groups and the control groups on individual support for the censorship apparatus and the regime in each of the two experiments as well as for the two studies combined.[15] As shown in Figure 1.3, the results are consistent across the two experiments. In particular, respondents exposed to censorship of both political and non-political content are reliably more supportive of the censorship apparatus (study 1: $\beta = 0.264$, $se = 0.082$; study 2: $\beta = 0.163$, $se = 0.042$; combined: $\beta = 0.185$, $se = 0.037$). Such results are robust to multiple hypotheses testing correction using the Benjamini-Hochberg method (see Appendix B.3) and indicate that the range of censorship targets matters for public reactions toward censorship. Additional

[15]Additional regression results with pre-treatment covariates are reported in Appendix B. The results are consistent with the main results presented in the main paper.

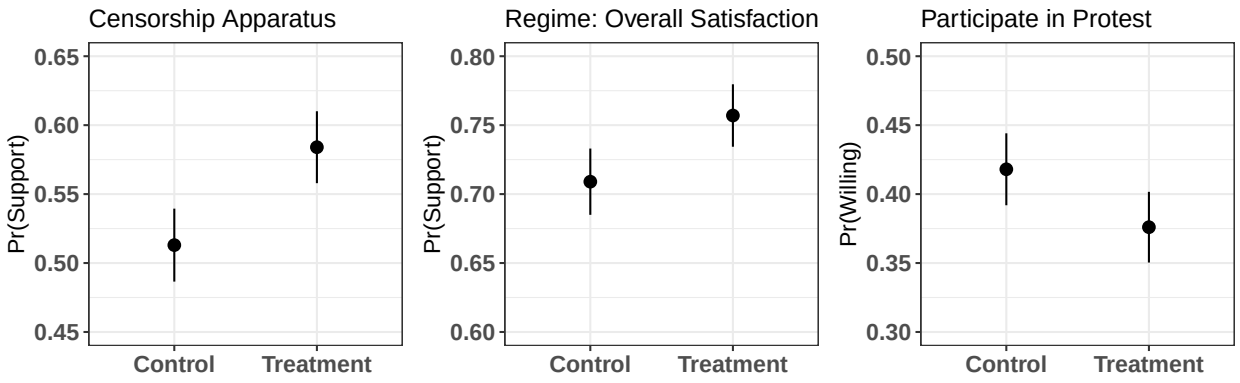
exposure to censorship of non-political content reduces respondents' backlash against the censorship apparatus even if they are also exposed to censorship of political content.



Notes: All outcome variables were measured on a five-point scale. The dots are the difference in means between the treatment and control groups. Bars indicate the corresponding 90% and 95% confidence intervals. *P*-values of all coefficients, except the willingness to protest in study 2, are still significant at the conventional level after the Benjamini-Hochberg correction.

Figure 1.3: Treatment Effects of Additional Censorship of Non-Political Content on Support for the Censorship Apparatus and the Regime (Difference-in-Means)

To further demonstrate the magnitude of the treatment effects, I translate the raw coefficients into proportions of respondents supporting the censorship apparatus. As shown in the left panel of Figure 1.4, combining the two studies, around 58.4% of the respondents in the treatment groups are strongly or somewhat supportive of the censorship apparatus, compared to only 51.3% of those in the control groups. Additional censorship of non-political content increases support by 7.1 percentage points.



Notes: The dots are the proportion of respondents who somewhat or strongly support censorship or the regime and are somewhat or definitely willing to participate in protests (two studies combined). The bars refer to the corresponding 95% confidence intervals.

Figure 1.4: Substantive Shifts in Support for the Censorship Apparatus and the Regime

For attitudes toward the regime, combining both studies, respondents exposed to censorship of both political and non-political content express higher overall satisfaction with the regime ($\beta = 0.133$, $se = 0.034$), higher support for both central ($\beta = 0.124$, $se = 0.034$) and local governments ($\beta = 0.141$, $se = 0.036$), as well as a lower willingness to protest ($\beta = -0.118$, $se = 0.046$). The effect sizes are smaller than the effects on support for censorship but still considerable given the fact that baseline support for the regime is already high in the control group. As shown in the middle panel of Figure 1.4, around 71% of the respondents are already supportive of the regime in the control group. As such, a 5 percentage point increase in regime support (76% among the treated) and a 5 percentage point decrease in willingness to protest (from 42% to 37%) are substantial in magnitude.

1.5.5 Discussion

An important issue I would like to address is whether the positive results are artifacts of preference falsification or social desirability bias (Kuran 1997), induced by the heightened censorship in the treatment group. To address this concern, Study 2 incorporates an

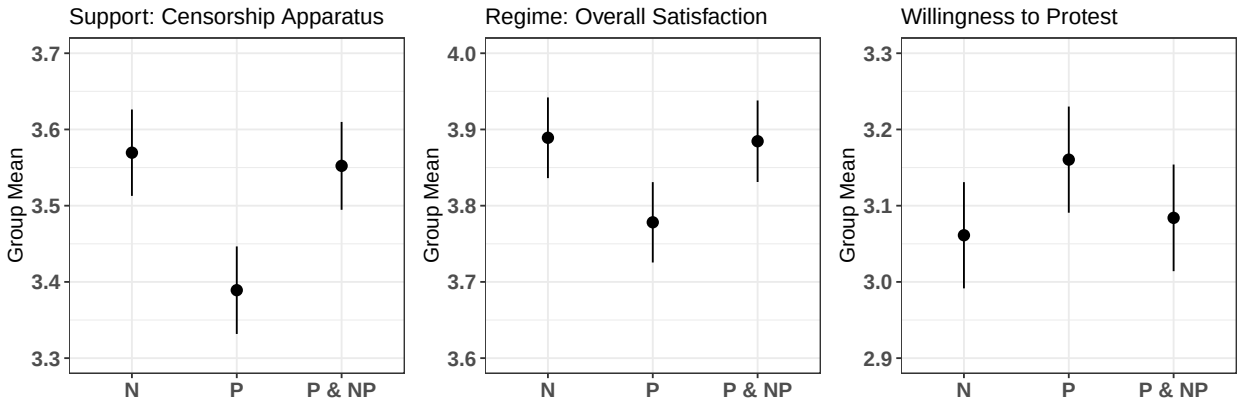
additional blank control group, where respondents are not exposed to censorship at all.[16] If the results in the main analyses are indeed driven by preference falsification induced by more censorship, we should expect the blank control group, without any censorship and therefore no priming of social desirability bias, to exhibit the lowest level of support. Moreover, we should expect censorship of political content to induce more severe preference falsification than non-political content.

However, as illustrated in Figure 1.5, these expectations do not hold. Compared to the blank control group without censorship (left bars in each panel), respondents exposed to censorship of political content (middle bars) reported lower support for the regime and its censorship apparatus, as well as a higher willingness to protest. These findings are consistent with existing studies that highlight the backlash effects stemming from political censorship (Pan and Siegel 2020; Roberts 2018), indicating that respondents are not coerced by the censorship primes into downplaying their aversion to censorship.

In contrast, when respondents are exposed to additional censorship of non-political content (right bars in each panel of Figure 1.5), their levels of support are indistinguishable from the blank control group without censorship (left bars). Such results demonstrate how additional censorship of non-political content normalizes public perception of the censorship apparatus as if the "new normal" is equivalent to the initial environment without censorship.

To further alleviate the concern that the treatment group might induce additional preference falsification compared with the control group, Study 2 also uses a list experiment to measure implicit support for censorship. The results show that overall, at most 7% of the respondents falsified their preference, and the treatment group has roughly the same

[16] Respondents in the blank control group still read the same ten social media snippets.



Notes: All outcome variables were measured on a five-point scale. The dots are the group means. Bars indicate the corresponding 95% confidence intervals.
N: No Censorship (Blank Control Group)
P: Censorship of Political Content Only (Control Group)
P & NP: Censorship of Both Political & Non-Political Content (Treatment Group)

Figure 1.5: Political Censorship Decreases Support for Censorship and the Regime, Increases Willingness to Protest, Whereas Additional Censorship of Non-Political Content Brings Support Back to Initial Levels Without Censorship

level of preference falsification as the control group (See Appendix B.4). Nevertheless, it is worth noting that fully isolating the impact of preference falsification might be difficult, and readers should still interpret these results with caution.

The experimental results have several implications. First, the results confirm that expanding the range of censorship targets beyond politically threatening content to seemingly harmless non-political content significantly increases public support for both the censorship apparatus and the regime. Such an expansion seems to be a plausible explanation of why most Chinese citizens are apathetic toward or supportive of government censorship. When exposed to censorship of non-political content, citizens update their beliefs about censorship and are more likely to view censorship as normal rather than repressive.

Second, the results show that repressive apparatuses like censorship might be popular and can increase regime support. This suggests that normalization is an independent channel

of authoritarian control, different from persuasion, which highlights the achievements of the government, and repression, which showcases regime strength and deters dissent.

1.6 Alternative Explanations and Limitations

Before concluding, let me also discuss an important alternative explanation. While the main experimental results demonstrate that censorship of non-political content leads to higher support, it might not be due to the normalization of censorship. Instead, individuals might simply favor censorship of non-political content, particularly when they believe the non-political content deserves to be censored. This is similar to public support for content moderation in democratic settings, such as removing misinformation, vulgar language, or pornography. Hence, bundling the censorship of political and non-political content increases the overall support for the censorship apparatus.

Although it is both possible and probable that citizens might like censoring some types of non-political content, it does not fully explain the phenomenon. As shown in the observational analysis, the range of non-political content is much broader than merely inappropriate content. In the experiment, I also validated the appropriateness of the non-political snippets and excluded those that were overtly absurd or inappropriate. More importantly, individuals who believe the censorship apparatus is normal should show higher support for the censorship of both political content and non-political content, whereas individuals who believe it to be political repression should show lower support for both. In contrast, the alternative bundling explanation predicts that support for censorship of political content remains constantly low and support for censorship of non-political content remains constantly high.

To obtain a better sense of whether the normalization theory explains the treatment effects above, in study 1, I asked respondents an additional question about whether they think "government control of the Internet is normal." I find individuals in the treatment group are significantly more likely to believe that censorship is normal ($\beta = 0.243$, $p = 0.002$), and this belief is a significant mediator between the treatment and support for censorship (ACME $= 0.149$, $p = 0.002$), directly supporting my normalization theory.

To further distinguish the normalization theory from the bundling argument, in study 2, I included two additional outcomes measuring support for the censorship of political content and non-political content respectively. First, comparing the blank control and control groups, I find that respondents in the control group with exposure to political censorship express significantly lower support for not only the censorship of political content ($\beta = -0.126$, $p = 0.004$) but also non-political content ($\beta = -0.095$, $p = 0.035$). Then, comparing the control and treatment groups, I find respondents in the treatment group express significantly higher support for the censorship of non-political content ($\beta = 0.101$, $p = 0.026$). However, albeit in the correct direction, the effect on political content ($\beta = 0.064$, $p = 0.155$) is not statistically significant at the conventional level. Moreover, I asked whether they think censorship targets political content, a proxy for the perceived level of repressiveness, and also find it to be a significant mediator both between the blank control and the control (ACME $= -0.017$, $p = 0.004$), as well as between the control and the treatment at the 0.1 level (ACME $= 0.011$, $p = 0.084$).

In summary, these findings demonstrate that support for non-political censorship is not consistently high, and support for political censorship is not consistently low either. Instead, it appears that specific measures of support are shaped by individuals' overarching perceptions of the censorship apparatus. Indeed, causal mediation analyses reveal that these overarching perceptions mediate the treatment effects. Nevertheless, the results do

not completely rule out the bundling argument, and it is possible that both explanations are not mutually exclusive. In essence, bundling effects could be occurring alongside normalization.

Apart from alternative explanations, there are also a few potential limitations to both the normalization theory and the empirical analyses. First, theoretically, does censoring *any* non-political content have the normalizing effect? It is well established in the political science literature that people care more about non-political issues like sports and entertainment than political issues (Carpini and Keeter 1996). Therefore, it is logical to argue that censoring some of the popular non-political topics might lead to backlash as well. Consistent with this logic, Hobbs and Roberts (2018) shows that the censorship of non-political content at a large scale, such as blocking certain platforms completely, is not without risk of public backlash. To avoid backlash, there may still be strategic considerations in choosing *which* non-political topics to censor. This remains an important topic for future research to explore.

It is worth noting that backlash against censorship can still occur, especially during critical moments marked by sudden increases in political suppression. For example, during the initial COVID outbreak in 2020 and the later COVID policy U-turn in 2022, the extensive suppression of political information did trigger public outrage, even in the presence of non-political censorship. This indicates that times of crisis can potentially undermine the effectiveness of normalization, precisely at the moments when the regime relies on it most.

Finally, my study relies on short-term, intense exposure to censorship to detect the normalizing effect which might not resemble the real world. Although previous research in psychology has shown that short-term, intense exposure has similar desensitizing effects compared with long-term exposure (Fanti et al. 2009), in reality, citizens are less likely to

encounter censorship as intensely as in the experiment and the actual normalization process might take longer. The fact that I found consistent results in both survey experiments two years apart, the first in 2020 and the second in 2022, alleviates part of the concerns. However, it might still be better to use longitudinal survey data to identify long-term normalization.

1.7 Conclusion

At the beginning of this study, I asked why existing literature that claims censorship awareness will lead to backlash cannot explain the reality that most Chinese citizens are either apathetic toward or supportive of the Chinese government's censorship policy. I pointed out that these existing studies primarily focus on censorship of government criticism and collective action, whereas the targets of censorship are much broader. Building on the desensitization theory in psychology, I argue that when the range of censorship is expanded beyond politically threatening content to seemingly harmless apolitical topics, citizens are more likely to view censorship as normal and less likely to react intensely and negatively.

The experimental and observational evidence that I presented supports my normalization theory. It shows the possibility that the normalization of censorship happens in China and the government attempts to create a benevolent image of the censorship apparatus. Moreover, expanding censorship targets has significant effects at the individual level. Citizens exposed to normalized censorship that targets both political and non-political content display significantly higher support for both the censorship apparatus and the regime.

The normalization theory can also be applied beyond China. Many authoritarian regimes employ similar strategies of diluting the repressive image of the censorship apparatus and desensitizing their citizens to information control. Moreover, autocracies are not the only systems of government that engage with censorship. Future research could extend the theoretical framework here to democracies and explore how regime type affects public reactions to and acceptance of censorship.

Besides censorship, normalization theory also provides an effective explanation of the phenomenon that some repressive policies in authoritarian regimes cause outcries and widespread attention in Western media whereas most people in the authoritarian regime do not have strong reactions to them. For example, Western media have widely reported on China's digital surveillance system powered by millions of digital cameras as well as China's social credit system (Kostka 2019). Yet Chinese citizens do not seem to be bothered much about all the surveillance. On one hand, the primary purpose of these surveillance systems is to prevent crime and ensure public safety, whereas repression of dissent only happens occasionally. This leads most Chinese citizens to believe that these surveillance systems are benevolent. On the other hand, Chinese citizens encounter these surveillance systems every day. Repeated exposure has effectively desensitized them, and these surveillance systems that cause outcries in the Western world are just part of normal life in China.

Chapter 2

Participatory Censorship in Authoritarian Regimes

2.1 Introduction

The conventional wisdom on government censorship in authoritarian regimes emphasizes its top-down nature, regarding it as a tool used by autocrats to silence criticism and prevent collective action that could destabilize the regime (Gueorguiev and Malesky 2019; King, Pan and Roberts 2013; Miller 2018; Pop-Eleches and Way 2021; Roberts 2018; Shadmehr and Bernhardt 2015). Given this understanding, existing research has extensively studied the ways in which ordinary citizens circumvent and resist government censorship activities (Chang et al. 2022; Chen and Yang 2019; Gläßel and Paula 2020; Han 2018; Hobbs and Roberts 2018; Pan and Siegel 2020; Roberts 2018, 2020).

However, contrary to conventional views, citizens in authoritarian regimes frequently participate in censorship by reporting online content. For example, in Turkey, Twitter users

systematically weaponize the report function against political opponents (Tufekci 2017). Similarly, in Russia, hundreds of users maliciously reported supporters of Ukraine on Facebook just before the Russian invasion of Ukraine (Nimmo and Agranovich 2022). In China, social media platforms explicitly ask their users to report each other for "violating Internet laws" (Jiang 2021) and reward those who report with "credit scores" (Cook 2019). Furthermore, surveys across the world demonstrate significant levels of popular support for government censorship, especially in authoritarian regimes like China (Dickson 2016; Wang and Mark 2015), Russia (Nisbet, Kamenchuk and Dal 2017), and the Middle Eastern monarchies (Martin, Martins and Wood 2016; Wike and Simmons 2015). *Why are many citizens in authoritarian regimes supportive of government censorship? Is it possible that participating in censorship increases citizens' support for it?*

This study takes a new step in this research field by exploring the consequences of a novel, bottom-up perspective of censorship. Instead of viewing censorship as a unilateral, top-down move by the government, I take a broader definition of censorship that includes any activities that contribute to the restriction of the public expression of or public access to information. I hypothesize that when ordinary users actively participate in the censorship process by flagging online content they disapprove of, their support for the censorship apparatus increases. This phenomenon could be attributed to several factors, including a reduction in the perceived responsibility of the government for censorship, the creation of cognitive dissonance, and individuals motivated to justify the censorship system they participated in.

Using an original online survey in China, I first provide a novel descriptive analysis of the prevalence of public participation in the censorship process in authoritarian regimes. I find that over half of the respondents self-reported having participated in censorship. The flagged content covered a wide range of topics. While inappropriate and socially

harmful content such as vulgar language is the most commonly flagged content, political discussions and tabloid gossip of entertainment stars are also widely reported. Additionally, public participation in censorship was prevalent across several demographic groups. Furthermore, I find a significant and positive correlation between participation and support for censorship.

To causally test the hypothesis, I conducted an original, pre-registered online experiment in a custom-engineered, simulated social media environment. The simulated social media page was not interactive, meaning each respondent completed the study independently. During the experiment, respondents in the treatment group were given an “encouragement" to participate in censorship by reporting the social media posts they encountered. I then used an instrumental variable analysis to estimate the complier average causal effect (CACE) of the experimentally induced censorship participation. Consistent with the theory, participation in censorship significantly increased individuals’ support for government censorship. I also find suggestive evidence that the institutional feature that allows public participation alone can increase support for censorship, alleviating concerns about experimenter demand effects and social desirability bias.

This study contributes to two streams of literature on authoritarian politics: government censorship and public participation. First, the study posits that censorship in authoritarian regimes should be perceived as a symbiotic relationship between the government and citizens, extending the literature that has examined authoritarian censorship solely from a top-down angle (Gallagher and Miller 2021; Gueorguiev and Malesky 2019; King, Pan and Roberts 2013; Lorentzen 2014; Miller 2018; Roberts 2018; Shadmehr and Bernhardt 2015). Moreover, the study provides novel empirical evidence of widespread public participation in the censorship process and explains how such participation shapes their opinions toward the government censorship apparatus. Importantly, the bottom-up perspective

helps to reconcile the empirical puzzle of why repressive apparatus such as censorship continues to garner significant popular support in authoritarian regimes (Dickson 2016; Wang and Mark 2015; Wike and Simmons 2015).

Second, the study extends the literature on public participation in authoritarian regimes by highlighting a novel and perhaps insidious form of public participation. While a wealth of literature has illustrated the causes and consequences of public participation in quasi-democratic institutions (Distelhorst and Hou 2017; Gandhi 2008; Gueorguiev 2021; He and Warren 2011; Manion 2015; Stromseth et al. 2017; Truex 2016, 2017) and contentious social movements (Fu and Distelhorst 2018), I show that public participation is also critical in the implementation of repressive policies like censorship. Encouraging citizens' participation in the censorship process consolidates public support for authoritarian regimes' repressive apparatus. This semblance of public participation ironically suppresses individual rights to free speech and contributes to the durability of authoritarian regimes in the Internet era.

2.2 Participatory Censorship: Bottom-Up Perspective

Censorship in authoritarian regimes has traditionally been viewed as a top-down process imposed by the state and social media platforms on ordinary citizens. For example, Roberts (2018) defines censorship as "the restriction of the public expression of or public access to information *by authority* [emphasis added]." Han (2018) also regards censorship as "tools used *by the state* [emphasis added] to limit the boundaries of online expression." Recent research has expanded the purposes of government censorship beyond silencing political opposition to rewarding regime supporters (Esberg 2020), but the consensus remains that censorship is a top-down tool of authoritarian regimes. However, this traditional view overlooks the role of ordinary users in censorship. This study takes a broader definition

of censorship that includes any activities that contribute to the restriction of the public expression of or public access to information and explores a novel perspective of censorship in which ordinary citizens participate by reporting online content.

Many authoritarian regimes have a history of public participation in repressive apparatus and political campaigns. During China's Cultural Revolution, for instance, ordinary citizens reported their friends, colleagues, and even families to the communist government as "counter-revolutionaries," which often led to brutal state repression (Dikötter 2016; Jiang 2021; Thurston 1984; Yang 2021). This phenomenon was also common in other dictatorships (Gregory 2009) and continues today with almost every social media platform having a function to report online content. While this feature can be used to flag inappropriate and harmful content, governments and citizens can and do abuse it for political gain (Nimmo and Agranovich 2022). For example, Tufekci (2017) documents that Turkish Twitter users have organized mass reporting of political opponents as spam to get their accounts suspended. In China, participation in censorship is especially prevalent. Official statistics claim to receive over 172 million censorship requests in 2022 alone (Xinhua 2023). Observers have even drawn parallels between this rising online "report culture" and the Cultural Revolution in the Mao era (BBC News 2020; Cook 2019; Jiang 2021).

Although similar reporting behaviors might occur in democracies as well, authoritarian regimes, such as China, are particularly keen on encouraging citizens to participate in the censorship process and report online content (Cook 2019; Jiang 2021). Since the establishment of the Central Leading Group for Cybersecurity and Informatization in 2014—an entity firmly under the control of Chinese leader Xi Jinping—it has effectively assumed control over the Cyberspace Administration of China (CAC). Subsequently, encouraging public participation in censorship has emerged as a pivotal objective for the CAC. Notably, the CAC houses a dedicated division called *jubao zhongxin*, exclusively

tasked with soliciting and addressing censorship appeals from ordinary users. The CAC even ran official propaganda campaigns at both central and local levels to promote public participation in censorship (CNS 2020; Sina 2023) and reward citizens for being “peer informants” (Cook 2019).

Why do authoritarian governments allow and even encourage such public participation? Theories of political participation in authoritarian regimes suggest that authoritarian governments allow public participation, including in censorship, to gather valuable information such as public policy preferences and potential social unrest due to the lack of democratic institutions (Distelhorst and Hou 2017; Gueorguiev 2021; Manion 2015; Stromseth et al. 2017; Truex 2016, 2017). In the context of censorship, authoritarian governments face the challenge of identifying messages that pose a threat to the regime (Gueorguiev and Malesky 2019; King, Pan and Roberts 2013). However, even in sophisticated regimes like China, complete control over the internet is difficult to achieve (Roberts 2018). To effectively censor a large amount of information, the Chinese government reportedly employs millions of censors (King, Pan and Roberts 2017), uses automated keyword filtering (Han 2018; Ng 2015), and employs a “friction" strategy to limit access to sensitive content (Roberts 2018).

To alleviate the information gathering problem, authoritarian governments also encourage ordinary users to participate in the censorship process (Cook 2019; Jiang 2021; Nimmo and Agranovich 2022; Tufekci 2017). The Chinese government, for instance, ran official propaganda campaigns in 2020 and 2023 to promote public participation in censorship (CNS 2020; Cook 2019; Sina 2023). User reports provide valuable signals for the government and social media firms to conduct censorship, reducing the cost of monitoring the internet. Platforms like Sina Weibo establish algorithms that consider user reports when determining the publicity and censorship of posts and accounts (Cook 2019; Jiang 2021). By mobilizing

millions of ordinary users to participate in censorship, the dynamic of censorship in China shifts from solely top-down control to a mixture of top-down control and bottom-up participation.

2.2.1 Participation and Public Support for Censorship

While previous studies have highlighted the negative effects of government censorship on public opinion towards the regime and its censorship apparatus (Gläßel and Paula 2020; Pan and Siegel 2020; Roberts 2018, 2020), this study proposes that when ordinary citizens participate in the censorship process by reporting online content, their support for the censorship apparatus actually increases. There are several reasons to believe that public participation may have this effect.

First, public participation in censorship can diffuse the government's responsibility and create a passive image of the government. By outsourcing repression to non-state actors, the regime can plausibly deny wrongdoing and evade political accountability (Ong 2022). In the case of censorship, the Chinese government mobilizes its pro-regime base to fabricate millions of posts to counter online critics (Chen and Xu 2017; Han 2015; King, Pan and Roberts 2017; Miller 2018), and outsources censorship to social media platforms (Han 2018; Miller 2018). These tactics divert blame for censorship away from the government and onto ordinary people and social media companies. Public participation in censorship further diffuses the government's responsibility. It creates the impression that the government is merely responding to public demand rather than initiating censorship (Luo and Li 2022; Zhao and Chen 2023). Therefore, even if citizens disapprove of certain censorship events, they may attribute them to other users' reports and are less likely to blame the government. Moreover, public participation creates the perception that censorship is not just the will of

the government but also the will of many ordinary users and that it is normal for online content to be removed.

Second, the cognitive dissonance theory, which posits that individuals experience psychological discomfort when their beliefs and behaviors are inconsistent (Festinger 1957), might also explain the effect of public participation in increasing support. Because there is an inconsistency between reporting online content and disliking censorship, such psychological discomfort motivates individuals to reduce the inconsistencies by changing their beliefs. Thus, despite initially opposing censorship, individuals may still participate in censorship to remove content they dislike, creating a cognitive dissonance that prompts them to justify their behavior and support government censorship to reduce psychological discomfort.

Similarly, the system justification theory suggests that individuals are motivated to defend and justify the social, economic, and political systems on which they depend (Jost 2020). In authoritarian regimes, this may lead individuals to justify initially unpleasant behaviors and experiences, such as censorship participation, to maintain a positive self-image and support for the existing regime. Furthermore, constant engagement in reporting online content may also contribute to the subconscious justification of such behavior and the censorship apparatus. Taken together, these theories support the notion that the bottom-up perspective of censorship may partly explain the popularity of the censorship apparatus in authoritarian regimes.

Hypothesis: As individuals participate more in the censorship process, they should display greater levels of support toward government censorship.

2.3 Study 1: Online Survey

Although participatory censorship is a widespread phenomenon in various autocracies (Tufekci 2017), this study focuses on China, one of the most sophisticated censorship regimes in the world, to illustrate the features and consequences of participatory censorship. To gauge the prevalence of public participation in censorship among the Chinese population and its correlation with support for censorship, I conducted an original online survey in December 2021.[17] The survey recruited 1,124 respondents through a Chinese online survey platform, who were then directed to an anonymous survey hosted by Qualtrics, a US-based survey platform. Respondents were selected using a quota sampling strategy to ensure a diverse range of socioeconomic backgrounds. However, like other online surveys in China (Huang 2018; Pan and Xu 2020), the sample may be younger and better educated than the general Internet population. To address this concern, the survey sample was weighted to resemble the Chinese Internet population in terms of gender, rural/urban location, region, age, and education.[18] In the main paper, I report results using the weighted sample and in the appendix, I report results using the original sample. The results are generally consistent.

2.3.1 Measurement

To measure *Participation in Censorship*, the survey asked respondents directly if they have ever reported online content or speech. Response options range from “never" to “multiple times per month," coded on a five-point scale. Although social desirability bias is a common concern in surveys conducted in authoritarian regimes, it is unlikely to be a significant issue in this study. As previously discussed, the Chinese government has expressed

[17]This study was approved by the Institutional Review Board (IRB) at the researcher’s home institution.
[18]For more information on the survey sample and weighting, see Online Appendix B.

support for such actions. As a result, participating in censorship is not a taboo in China and respondents are unlikely to fear reporting their past behaviors had they participated. Moreover, respondents are also unlikely to falsely report having participated before if they have never because there has not been any social norm to participate. For those participants who report having participated in censorship, they are asked to specify the type of content they requested to be removed from a list of options. This list includes political content, such as political news, commentary, opinions, rumors, and foreign media coverage of China. Additionally, non-political content like entertainment and advertisements are also included, as well as inappropriate content such as vulgar language and pornography.

The main outcome variable in this analysis is *Support for Censorship*. Respondents were asked whether they agreed with the statement that the government should actively control the Internet and remove content that it deems inappropriate. Additionally, I measured their *Support for Censorship of Political Content* and *Support for Censorship of Non-Political Content* by asking whether the government should control online discussion of government policies and party leadership, as well as entertainment stars and popular culture. All variables were measured on a five-point scale.

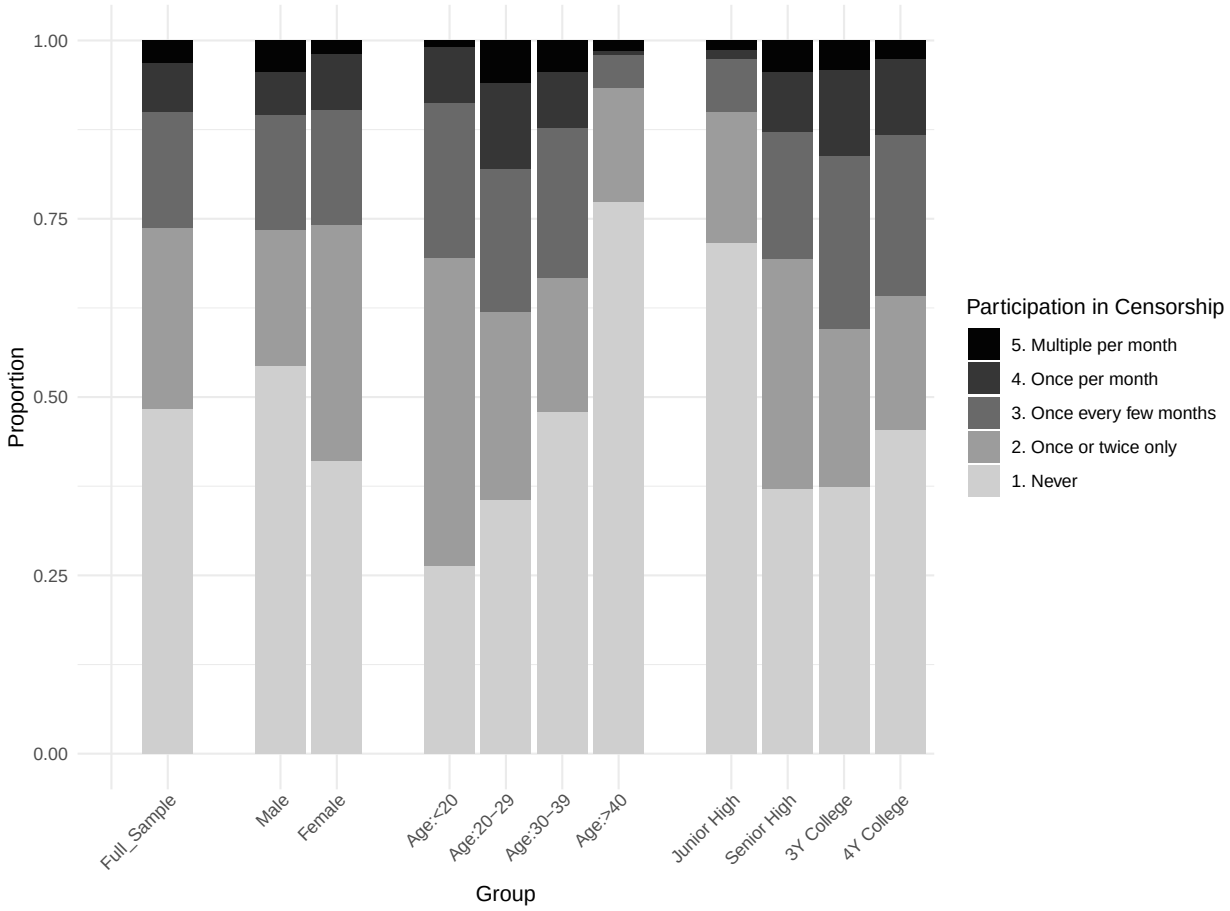
Two sets of control variables were included: demographic covariates and predisposition covariates. Demographic covariates included *education*, *age*, *gender*, and *urban/rural location*. Predisposition covariates included *party membership*, *political interests*, *political ideology*, and *economic ideology*. The first two covariates are commonly used in surveys in China (Huang 2018), while the third and fourth questions measuring ideology were borrowed from Pan and Xu (2020).

2.3.2 Results: Prevalence of Public Participation

How prevalent is public participation in censorship among the Chinese public? Figure 2.1 presents the distribution of self-reported participation in censorship using the weighted sample. As shown in the first bar on the left, more than 50% of the weighted sample report having previously participated in censorship. More than 25% have participated at least once every few months. These results demonstrate that public participation in censorship is prevalent among Chinese internet users. Such behavior is especially common among younger generations who are regularly exposed to online discussions about the use of reporting as a strategy to censor opposing opinions (Luo and Li 2022; Zhao and Chen 2023). Almost three-quarters of respondents under 20 years old report having such censorship experience, and around two-thirds of respondents in their 20s report similar experiences. For these young people on the Internet, flagging online content is both common and normal.

What specific content did these “participating respondents" report? The vast majority of participating respondents (around 90%) reported inappropriate content online, including pornography and vulgar language. A significant proportion of them also reported political content, with about 50% of participating respondents (or 25% of all respondents) reporting having flagged political content. Younger and better-educated respondents were more likely to participate in political censorship, regardless of their political predisposition. Finally, around one-third of participating respondents reported censorship of entertainment and cultural content. In summary, the descriptive analysis indicates that there are significant levels of public participation in censorship across different demographic groups and categories of online content.[19]

[19] see Online Appendix C for more descriptive analyses of the reported content.



Note: All observations are weighted by gender, rural/urban location, region, age group, and education. The unweighted sample shows similar patterns (see Online Appendix C).

Figure 2.1: Distribution of Self-Report Participation in Censorship

2.3.3 Participation and Censorship Support

Do individuals with higher levels of participation in censorship hold more favorable views toward the censorship apparatus? Table 2.1 reports the results of OLS models that investigate the relationship between participation in censorship and support for the censorship apparatus among the Chinese public. Consistent with the hypothesis, individuals who have more actively participated in the censorship process are more likely to support government censorship. Specifically, for each additional level of participation in

censorship, there is a .084 increase in support for censorship on a five-point scale, even after controlling for demographic and predisposition covariates. This effect size is equivalent to 8.34% of a standard deviation of the dependent variable, which suggests that participation in censorship is a significant predictor of support for censorship. As robustness checks, I first use alternative modeling strategies such as ordered logistic regression models. I also re-coded participation as a binary variable indicating whether the respondent has ever participated or not. The results are consistent with the main analyses.[20] The regression analysis results suggest that individuals who have flagged more content in the past are more likely to believe that the government should actively remove content it deems harmful or inappropriate.

Table 2.1: Correlation between Participation in Censorship and Support for Censorship

	Support for Censorship		Support for Censorship of Political Content		Support for Censorship of Non-Political Content	
	(1)	(2)	(3)	(4)	(5)	(6)
Participation	0.099***	0.084***	0.085***	0.105***	0.007	0.027
	(0.028)	(0.028)	(0.030)	(0.030)	(0.033)	(0.033)
Constant	2.616***	2.067***	2.549***	2.583***	2.367***	2.387***
	(0.131)	(0.182)	(0.136)	(0.191)	(0.150)	(0.211)
Demographic	✓	✓	✓	✓	✓	✓
Predisposition		✓		✓		✓
Weighted	✓	✓	✓	✓	✓	✓
N	1,088	1,071	1,084	1,066	1,086	1,068
Adjusted R^2	0.048	0.106	0.046	0.074	0.034	0.070

Notes: Dependent variables are indicated in column headings and are measured on a five-point scale. Standard errors are in parentheses.
$^{*}p < .1$; $^{**}p < .05$; $^{***}p < .01$

The correlation between public participation in censorship and support for government censorship also holds for political content, as evidenced by the results in columns 3 and 4

[20] See Online Appendix D.2.

of Table 2.1. Respondents who have engaged more in censorship are more likely to endorse the government's active control of political news and discussions, and the magnitude of the effect is even stronger than that of the main models analyzing general support for censorship. However, there is no significant correlation between censorship participation and support for regulating non-political content. This could be due to the relatively uncontroversial nature of government regulation of entertainment content and popular culture, resulting in less variation in the dependent variable.

Given the overall significant correlation between participation and support for censorship, an important theoretical question arises: is this correlation primarily due to the censorship behavior itself, regardless of the content being reported, or is it contingent on the content being reported? Table 2.2 demonstrates the correlation between specific types of participation, such as reporting political content versus non-political content, and support for censorship in general, as well as censorship of each content type.

As shown in columns 1 and 2 of Table 2.2, higher levels of participation in political censorship significantly correlate with higher support for censorship. Likewise, reporting entertainment and cultural content also exhibits a positive correlation with support, albeit with a lesser degree of significance ($p = 0.138$). Moreover, both types of reporting behaviors are significantly correlated with support for censorship of political content, as indicated in columns 3 and 4. These results suggest that the reporting behavior itself contributes to higher acceptance of the censorship apparatus. Even reporting non-political content might have a significant spillover effect and increase support for political censorship.

In summary, the observational analyses provide evidence for the argument that public participation in censorship is widespread among the Chinese population, with approximately 50% of the respondents self-reported having engaged in censorship. Furthermore,

Table 2.2: Correlation between Specific Types of Participation and Support for Censorship

	Support for Censorship		Support for Censorship of Political Content		Support for Censorship of Non-Political Content	
	(1)	(2)	(3)	(4)	(5)	(6)
Participation (Political)	0.059**		0.121***		0.004	
	(0.030)		(0.032)		(0.035)	
Participation (NonPolitical)		0.053		0.079**		−0.001
		(0.036)		(0.037)		(0.043)
Constant	2.151***	2.159***	2.618***	2.682***	2.435***	2.443***
	(0.178)	(0.181)	(0.187)	(0.190)	(0.207)	(0.211)
All Covariates	✓	✓	✓	✓	✓	✓
Weighted	✓	✓	✓	✓	✓	✓
N	1,071	1,071	1,066	1,066	1,068	1,068
Adjusted R^2	0.102	0.100	0.076	0.067	0.069	0.069

Notes: Dependent variables are indicated in column headings and are measured on a five-point scale. Standard errors are in parentheses. The independent variables are participation in political censorship and participation in censorship of entertainment and cultural content. Both demographic and predisposition covariates are included in all models.
*p < .1; **p < .05; ***p < .01

the analysis demonstrates a positive association between participation in censorship and support for censorship, which is consistent with the main hypothesis. These findings have significant implications for understanding censorship dynamics in authoritarian regimes. By delegating some censorship power to ordinary citizens, the Chinese government may have successfully increased public support for the censorship apparatus, contributing to the regime's stability and control over online discourse. These findings highlight the relevance of the bottom-up perspective of censorship in authoritarian regimes. A significant proportion of Chinese internet users engaged in flagging content highlights the importance of analyzing the implications of such participation to develop a comprehensive understanding of how citizens in authoritarian regimes perceive government censorship and repressive policies more broadly.

2.3.4 Mechanisms and Robustness

In the theory section, I presented three potential reasons to explain how public participation in censorship might increase support for the censorship apparatus. To examine whether participation reduces the government's responsibility for censorship, the survey asked respondents about who they believed should be accountable for censored content: netizens, the government, or platforms. Using OLS regression models with all relevant covariates and adjusted by sample weights, I find higher levels of participation in censorship are indeed significantly and negatively associated with perceived government responsibility ($\beta = -0.112$, $p = 0.023$). Similarly, causal mediation analysis demonstrates that perceived government responsibility has a significant and negative causal mediation effect on the relationship between participation and support for censorship.

It is more challenging to directly test the remaining two mechanisms — the cognitive dissonance theory and the system justification theory. Nonetheless, both theories imply that participation in censorship would have a more significant impact on increasing support for the censorship apparatus than increasing support for the authoritarian regime as a whole. To explore this possibility, the survey examined respondents' overall satisfaction with the Chinese regime, as well as their evaluation of the government's performance. Using OLS models, I indeed find no statistically significant relationships between participation in censorship and regime support ($\beta = -0.012$, $p = 0.632$). As a result, the findings suggest that participation has a stronger impact on censorship support than regime support, which provides suggestive evidence for both the cognitive dissonance theory and the system justification theory. In summary, all three mechanism arguments appear plausible for explaining the positive effect of public participation on support for censorship.

However, the observational analyses above may be susceptible to the possibility of omitted variable bias or reverse causality. To mitigate these concerns, I conducted a sensitivity analysis following Cinelli and Hazlett (2020) to test the robustness of the main model (column 2 in Table 2.1) to potential unobserved confounders or reverse causal relationships. The results of the analysis, shown in column 5 of Table 2.3 (*RV*), suggest that a potential unobserved confounder or a reverse causal arrow through such a confounder would need to explain at least 8.8% of the residual variance of both the treatment and the outcome to explain away the estimated treatment effect. In comparison, an unobserved confounder as strong as *Economic Ideology*, the most significant predictor in the main model, can only explain 4.2% of the residual variance. Therefore, the model is at least robust to an omitted variable as strong as the most significant covariate in the current model.

Table 2.3: Sensitivity of the Main Regression Model to Unobserved Confounders

Treatment:	Outcome: *Support for Censorship*					
	Estimate	Std. Error	t-value	$R^2_{Y\sim D\|\mathbf{X}}$	RV	$RV_{\alpha=0.05}$
Participation in Censorship	0.084	0.028	2.997	0.8%	8.8%	3.1%
df = 1061; *Bound (Z as strong as Economic Ideology)*: $R^2_{Y\sim Z\|\mathbf{X},D}$ = 4.2%, $R^2_{D\sim Z\|\mathbf{X}}$ = 0.4%						

Notes: RV stands for robustness value, the proportion of residual variance of both treatment and outcome a confounder needs to explain in order to explain away the treatment effect. $RV_{\alpha=0.05}$ is the RV such that the treatment is no longer statistically significant at 0.05 level. $R^2_{Y\sim D|\mathbf{X}}$ is the proportion of residual variance of treatment that a confounder needs to explain in the extreme scenario that it explains 100% of residual variance of the outcome Cinelli and Hazlett (2020). The benchmark is *Economic Ideology*, the most significant predictor in the main model.

It should be noted that this study does not claim that participation is exogenous to individuals' political predispositions and prior support for censorship. Although I acknowledge the possibility that more pro-censorship citizens might be more likely to participate, the

goal of this study is to explore the downstream effects of political participation in repressive apparatus, such as censorship, on individuals' support for it. To address concerns about the correlation being driven by a reverse causal relationship and to causally test the hypothesis, I conduct an experiment that randomly manipulated individuals' participation in censorship. The following section describes the design and results of the experiment.

2.4 Study 2: Survey Experiment

Building on the first study, I conduct an original, pre-registered online survey experiment using a custom-engineered, simulated social media page.[21] The simulated social media environment was not interactive, meaning each participant still completed the survey independently. The experiment aims to test the causal effects of participation in censorship on public support for government censorship. Since it is challenging to manipulate censorship behavior directly, I employed an instrumental variable approach, where I provided respondents in the two treatment groups with options and encouragement to report simulated social media posts. I then measured the complier average causal effect (CACE) on support for censorship among respondents who actually participated in reporting the simulated social media posts (Aronow and Carnegie 2013; Marbach and Hangartner 2020). This approach allows me to estimate the causal effects of censorship participation induced by the experimental treatments. Additionally, I also tested the intention-to-treat effect of the treatments by comparing the group means of the treatment groups to the control group.

Conducting an experiment in a simulated social media environment offers several advantages over a similar field experiment on real social media. First, it avoids ethical concerns

[21]This study was approved by the Institutional Review Board (IRB) at the researcher's home institution and was pre-registered on Open Science Framework.

associated with field experiments, particularly given the current political climate in China, which is hostile to political research. Such an experiment might put both participants and researchers at higher risk of authoritarian repression. Furthermore, encouraging respondents to participate in censorship in the real world is normatively undesirable and might further contribute to the reporting culture on the Chinese Internet. Conducting the experiment in a simulated setting limits the potential negative impact of the research. Finally, reporting behaviors are usually not publicly observable in the real world, and in my experiment, a significant proportion of respondents object to the idea of reporting and refuse to participate in censorship even after encouragement. An experiment embedded in a simulated environment enables me to measure participation, estimate treatment effects among those who comply with the treatments, and identify those who refuse to participate regardless of treatments.

2.4.1 Procedure

The survey experiment was conducted in June 2022 in China. Similar to the online survey in study 1, I recruited around 4,000 respondents from a Chinese online survey platform and then directed them to Qualtrics, where they completed the survey anonymously. As was the case with study 1, the sample covers a wide range of socioeconomic backgrounds but is younger and better educated than the general Internet population.

The experiment consists of three parts. First, I measured pre-treatment covariates. Second, I randomly assigned respondents to one of three groups, including a control group and two treatment groups. All participants were asked to use the simulated social media page where they read ten posts related to a hotly debated current event in China in 2022, the

Xuzhou chained woman incident.[22] The ten posts were adapted from actual Sina Weibo posts with modified user names and avatars. Five of the posts are pro-government or nationalistic, while the remaining five are anti-government or pro-individual rights. The order of the posts was randomized. After reading and potentially reacting to the posts, I measured respondents' support for censorship.

On the simulated social media page, I built multiple buttons that the respondents can click under each post. In the control group, these buttons are: "Like," "Share," and "Comment." In both treatment groups 1 and 2, I built an additional "Report" button under each post that allows respondents to flag the post, mimicking the real-world participation process in censorship. Furthermore, in treatment group 2, respondents received an additional "encouragement" to use the "Report" button. Specifically, respondents in treatment group 2 were shown the following paragraph:

> *We are especially interested in what posts you want to report. Please choose at least two posts that you think should be removed by the Internet regulator, and press the Report button to let us know.*[23]

Table 2.4 summarizes the experimental design and treatment assignments. By comparing the difference in means between the control group and treatment group 1, we can test the impact of the institutional feature that permits public participation in censorship. Although this is not a direct test of the main hypothesis regarding the effects of participation behavior, it offers additional insights into the bottom-up perspective of authoritarian

[22] In January 2022, a video of a trafficked woman held in chains in a hut in Fengxian County, Xuzhou City for years went viral. Government officials were heavily criticized on social media for causing such tragedy and, more importantly, trying to cover it up. However, the incident coincided with the 2022 Beijing Winter Olympics. Such timing prompted many patriotic regime supporters to argue that this was a conspiracy to defame China.

[23] Although respondents were encouraged to report at least two posts, they were not forced to do so. Non-compliance was allowed.

censorship. If we observe significant effects from merely providing the institutional feature for public participation, it may suggest that the effect sizes would be even more substantial if respondents actually participated in the censorship process. Furthermore, having both treatments mitigates the concerns of experimenter demand effects and social desirability bias. Since respondents in treatment group 1 were not explicitly instructed to participate in censorship, they were less likely to infer the objectives of the experiment or the social norms around censorship participation, reducing the likelihood of changing their behavior and providing socially desirable answers.

Table 2.4: Summary of Experimental Design and Treatments

Groups	Control	Treatment 1	Treatment 2
Buttons Under Simulated Posts	Like, Share, Comment	Like, Share, Comment, **Report**	Like, Share, Comment, **Report**
Encouragement Message	No	No	Yes

To measure respondents' participation, I use a binary variable indicating whether they clicked any of the "Report" buttons. As in the previous study, the main dependent variable is *Support for Censorship*, and I also measure *Support for Censorship of Political Content* and *Support for Censorship of Non-political Content*. To ensure balance across the three experimental groups, I include ten pre-treatment covariates in the analysis. Among these covariates, four are demographic variables, including *education*, *age*, *gender*, and *region*, which are commonly used in experiments across various contexts. The remaining six covariates are predisposition covariates, including *party membership*, *nationalism*, *political interests*, *ideology*, *social media usage*, and *foreign connection*. All covariates are balanced across the three experimental groups.[24]

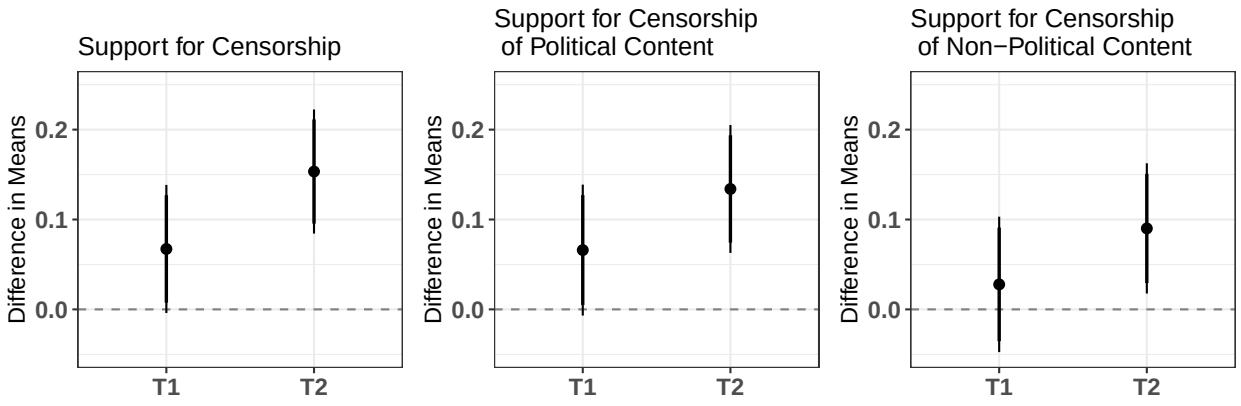
[24]See Online Appendix E.3.

2.4.2 Results: Difference in Means

I first report the overall results of the experiment by comparing the group means of the outcome variables across the three groups and then elaborate on the instrumental variable analysis that more precisely identifies the causal effect of participation on support for censorship. In the control group, none of the respondents reported any simulated posts as they were not allowed to do so. In treatment group 1, 43% of the respondents clicked the "Report" buttons on the simulated social media page, while 64% of the respondents in treatment group 2 did so. Thus, both treatments successfully induced participation among respondents who would not have otherwise participated.

Figure 2.2 presents the difference in means using the control group as the reference group. The left bar in each panel compares the control group and treatment group 1, and the results suggest that respondents who were given the opportunity to report simulated posts expressed higher support for censorship in general ($\beta = 0.067$, $p = 0.065$) and censorship of political content ($\beta = 0.066$, $p = 0.076$). Both results are statistically significant at the 0.1 level. However, I did not find a significant difference in support for censorship of non-political content, possibly due to the political nature of the selected topic. These findings suggest that simply providing the institutional feature to flag online content, without explicit messages that might induce experimenter demand effects or social desirability bias, can potentially increase individuals' support for censorship.

Moving on to treatment group 2, the right bars in each panel demonstrate that respondents who were given the option to report and encouraged to flag online content showed significantly higher levels of support for government censorship in general ($\beta = 0.153$, $p < 0.001$), as well as support for censorship of political content ($\beta = 0.134$, $p < 0.001$) and non-political content ($\beta = 0.090$, $p = 0.015$) in particular. The effect sizes are larger than



Note: All three outcome items are measured on a five-point scale. The control groups are the reference groups and bars indicate 90% and 95% confidence intervals.

Figure 2.2: Difference in Means of the Outcome Variables

treatment group 1, indicating that the additional participation in censorship induced by the encouragement message further increases support for the censorship apparatus.

All in all, the overall results of the experiment provide strong and consistent evidence for the central argument that increased participation in censorship leads to higher support for the censorship apparatus. The institutional feature that allows public participation and the encouragement treatment that directly increases reporting behaviors both generate significant support for censorship, giving us greater confidence that the increase in support is not merely an experimenter demand effect or due to social desirability bias. However, it is important to note that neither treatment directly measures reporting behavior. To estimate more precisely the effect of reporting behaviors on support for censorship, the next section introduces the instrumental variable analysis and reports its findings.

2.4.3 Instrumental Variable Analysis

To directly identify the effect of participation in censorship, I use both treatments as instruments to estimate the complier average causal effect (CACE) of participation in censorship (clicking the report button) on support for censorship. Formally:

$$\begin{aligned} \text{Clicking the Report Button}_i &= \alpha + \gamma \cdot \text{Treatment Group} + \lambda Z_i + \epsilon_i \\ Y_i &= \xi + \beta \cdot \widehat{\text{Clicking the Report Button}_i} + \delta Z_i + \mu_i \end{aligned}$$

where Y_i is the outcome measure; Z_i is a vector of pre-treatment covariates; and β is the CACE. In the instrumental variable analysis, I included all three groups and treated the two treatment conditions as factors that independently influence participation in censorship. In Online Appendix F.2, I present several alternative instrumental variable models, including (1) using an ordinal variable for the two levels of treatment and (2) analyzing the data with only two of the three groups. The results from alternative modeling strategies do not alter the substantive interpretation of the main results.

Table B.12 reports the results from instrumental variable analyses and the CACE of participating in censorship in the simulated social media environment. Consider, first, column 2. After controlling for pre-treatment covariates, participation induced by the encouragement treatment significantly increases respondents' general support for government censorship ($\beta = 0.219$, $p < 0.001$). This again provides direct support for the main hypothesis and more importantly, it addresses the concerns in Study 1 that the causal arrow might be reversed. The magnitude of the treatment effect on support for censorship is considerable, equivalent to 25% of a standard deviation. This is a substantial increase given that the baseline support for censorship is already high in the control group. The instrumental variable analysis also indicates that censorship participation induced by both treatments increases

specific support for censorship of political content ($\beta = 0.199$, $p < 0.001$) and non-political content ($\beta = 0.124$, $p = 0.028$), further supporting the theoretical expectations.

Table 2.5: Complier Average Causal Effects (CACE) of Participating in Censorship on Support for Censorship

	Support for Censorship		Support for Censorship of Political Content		Support for Censorship of Non-Political Content	
	(1)	(2)	(3)	(4)	(5)	(6)
Report Click	0.228***	0.219***	0.202***	0.199***	0.130**	0.124**
	(0.054)	(0.052)	(0.055)	(0.054)	(0.058)	(0.056)
Constant	3.484***	1.954***	3.463***	2.082***	3.673***	2.386***
	(0.024)	(0.110)	(0.025)	(0.114)	(0.025)	(0.119)
Covariates		✓		✓		✓
N	3,990	3,764	3,997	3,770	3,989	3,763

Notes: Dependent variables are indicated in column headings and are measured on a five-point Likert scale. Standard errors in parentheses.
Report click is a binary variable indicating whether the respondents have clicked any of the "Report" buttons on the simulated social media page.
*p < .1; **p < .05; ***p < .01

To check the robustness of the treatment effect, I employ an alternative measure of participation in censorship and re-run the instrumental variable analyses.[25] Instead of a binary variable indicating whether the respondents clicked any of the "Report" buttons, I use the count of times respondents clicked the "Report" button. The results from this analysis are consistent with the main analyses, indicating that the additional clicking of the "Report" button induced by the encouragement treatment leads to a significant increase in support for censorship. Additionally, I have subset the data to estimate the effect of participation induced by the option to report and the effect of participation induced by the encouragement message separately. Both analyses reveal a significant CACE of reporting simulated posts on the support of censorship.

[25]See Online Appendix F.2.

While the instrumental variable analyses indicate significant and positive treatment effects of censorship participation on support for government censorship, it is important to compare the background characteristics of compliers, those who participated when given the opportunity, and non-compliers, those who did not even when the opportunity was present, to comprehend which demographic groups drive the results (Marbach and Hangartner 2020). In treatment group 1, individuals who clicked the “Report" buttons tended to be younger and more familiar with social media. In treatment group 2, the additional participants due to the encouragement message were also young, well-educated, less nationalistic, and had more foreign connections. In contrast, the non-compliers, those who never click the “Report” buttons, tended to be older, nationalists with few foreign connections and limited social media exposure. Therefore, the results of the instrumental analyses are primarily driven by younger generations who tend to be more educated and familiar with social media. Such findings are somewhat consistent with Study 1 where the age group appears to be the primary determinant of participation (See Figure 2.1). Thus, it is less certain whether there are comparable effects of participation on support for censorship among older and more nationalistic individuals. However, the composition of compliers and non-compliers may differ depending on the chosen topic and posts, and the profiling outcomes should be interpreted with caution. Future research should test the hypothesis in various contexts to improve the generalizability of the findings.

2.4.4 Implications

To sum up the findings in the experiment, both the institutional feature that allows public participation in censorship and the message that directly encourages participation increase support for censorship to various degrees. The instrumental variable analysis further

demonstrates that the treatment effects are large and significant among respondents who actually participated in reporting the simulated social media posts.

The implications of the experimental results are three-fold. First, they echo the observations made in the previous survey that public participation in the censorship process is a widespread phenomenon in authoritarian China. As demonstrated by the treatment groups, approximately 40% of the respondents engaged in censorship when given the chance, and a simple encouragement message resulted in nearly two-thirds of the respondents participating in censorship.

Second, the results reaffirm the positive relationship between widespread participation in the censorship process and public support for censorship, highlighting the importance of a comprehensive understanding of authoritarian censorship that combines top-down control and bottom-up participation. The institutional feature that permits public participation and the encouragement message that directly increases reporting behaviors can both generate significant support for the censorship apparatus. Furthermore, the results are unlikely to be merely an experimenter demand effect since respondents in treatment group 1 expressed higher support even without explicit instructions to flag online content.

Third, the profiling of compliers indicates that the effect is mainly driven by young, well-educated individuals with foreign connections. This group is traditionally perceived as less vulnerable to authoritarian controls and more receptive to Western values (Huang 2015). The findings suggest that authoritarian regimes can leverage public participation in repressive apparatus such as censorship as a means to generate support for its repressive apparatus among this young and educated demographic, thereby maintaining popular support across generations. Overall, the experiment sheds light on the symbiotic relationship between the authoritarian government and its citizens in the censorship process,

particularly the bottom-up perspective that existing studies tend to overlook, and provides causal evidence for the positive effects of public participation on support for repressive apparatus such as censorship.

2.5 Conclusion

Repressive institutions such as censorship have long been viewed as instruments of top-down suppression employed by authoritarian regimes to quash political opposition. This study offers a new perspective by examining the prevalence and significance of bottom-up public participation in censorship in authoritarian regimes. Through two original online surveys conducted in China, one observational and one experimental, I demonstrate that public participation in censorship is a pervasive phenomenon that significantly shapes public opinion towards the censorship apparatus. Therefore, censorship in authoritarian regimes should be viewed as a symbiotic relationship between the government and its citizens, echoing the longstanding history of public involvement in repressive apparatus and political campaigns, such as the Cultural Revolution.

The findings in this study highlight the discrepancy between the common understanding of repressive authoritarian apparatus, such as censorship, in the Western world and how ordinary citizens in authoritarian regimes perceive and interact with these repressive apparatuses. For many citizens in China, censorship and other repressive institutions have been normalized as part of the rules of political life. Rather than fighting against the rules, individuals in these regimes take advantage of the censorship apparatus to suppress opposing views (Luo and Li 2022; Tufekci 2017; Zhao and Chen 2023). Consequently, citizens no longer view the regime as the oppressor, but rather as a powerful arbitrator

of censorship demands that they must win over in their internal struggles against fellow citizens.

Beyond autocracies, censorship has become an important social issue in many democracies including the United States, and public participation in censorship has also become more prevalent. Platforms such as Twitter have introduced community-based bottom-up content moderation projects that involve public participation. Although democracies might care about different policy implications, such as electoral integrity, compared with their authoritarian counterparts, it is still important to examine the consequences of these content moderation projects, because the balancing act of fighting against misinformation and preserving freedom of speech is difficult yet critical for sustaining a healthy democracy.

Chapter 3

How Chinese Censorship Allows Public Discourse on Democracies but Not Their Institutions

3.1 Introduction

The threat posed by information about foreign liberal democratic regimes has been a persistent concern for authoritarian leaders throughout history (Huang 2015). As far back as the French Revolution, European monarchies feared that their citizens gaining knowledge and engaging in discussions about the French Revolution could incite rebellion within their own territories. More recently, events such as the third wave of democratization and the Arab Spring have demonstrated the significant domestic impact that public awareness of foreign democratization events can have (Gläßel and Paula 2020; Gleditsch and Ward 2006; Huntington 1991; Steinert-Threlkeld 2017). Given the significance of international

influence on authoritarian survival, *how do authoritarian regimes manipulate information about foreign liberal democracies*?

Conventional wisdom, namely state critique theory, suggests that censorship targets content critical of the regime and the authoritarian system (Dukalskis 2021; Esberg 2020; Gueorguiev and Malesky 2019; Han 2018; King, Pan and Roberts 2013; Lorentzen 2014). This implies that positive coverage of liberal democracies will be heavily censored while negative propaganda about liberal democratic regimes will flood the public discourse (Carter and Carter 2023; Deng 2023; Gläßel and Paula 2020; Rozenas and Stukal 2019; Mattingly and Yao 2022). However, not every positive news about democracies threatens authoritarian survival, particularly during a period when many autocracies' economies are developing rapidly (Luo and Przeworski 2019). Moreover, due to the growing tension between democratic and autocratic powers, such as between the US and China, the Chinese social media landscape is already hostile toward foreign rivals like the US and Japan,[26] decreasing the need to suppress positive coverage of democracies.

In this study, I propose a more nuanced theory to explain the authoritarian censorship of foreign liberal democracies. Specifically, I theorize that the primary objective of censorship is to keep the public uninformed or misinformed about democratic institutions and processes. This is because citizens in authoritarian regimes have little experience or prior knowledge about democratic institutions and processes, giving authoritarian governments an advantage in manipulating the public to believe themselves as "democratic" and setting up nominally democratic institutions to consolidate autocratic rules (Gandhi 2008; Svolik 2012). Consequently, online discussions about democratic institutions and processes abroad, even negative ones, become especially dangerous to autocrats as they expose democratic practices different from the distorted domestic ones to the public. Moreover,

[26]As shown below, over 70 percent of the pre-censorship content is negative toward democratic regimes.

the generally low level of political knowledge regarding democratic institutions makes censorship less likely to backfire, compared to censoring socioeconomic conditions, which the public can benchmark on their own experiences (Geddes and Zaller 1989). Taken together, authoritarian regimes are likely to employ stricter censorship on public discourse concerning *democratic institutions*, such as elections, legislative debates, judicial reviews, checks and balances, free media, and liberal democratic values more broadly.

To test this hypothesis, I provide the first systematic analysis of how authoritarian regimes censor domestic discussion about liberal democracies. I focus on the case of China, one of the most influential authoritarian regimes that fiercely competes with liberal democracies, especially the United States, and strives to be the new global leader in many aspects. China also provides an ideal case for this study as its censorship apparatus is among the most sophisticated in the world (King, Pan and Roberts 2013). Using around 100,000 articles about liberal democracies on WeChat, the largest social media platform in China, between 2018 and 2022, I show that content involving democratic institutions (30 percent censored) is three times more likely to be censored than content about socioeconomic conditions and policy outcomes in liberal democracies (10 percent censored). Such a higher level of censorship persists, even when the stance is negative toward democracies, depicting gridlock, inefficiencies, or even shutdowns of democratic institutions. In contrast, the effect of stance and sentiment on censorship is marginal.[27]

My findings make significant contributions to our knowledge about the role of censorship and institutions in authoritarian politics. Over the last decade, a burgeoning body of literature has unveiled the multifaceted nature of authoritarian censorship, including the suppression of political criticism (Golovchenko 2022; Gueorguiev and Malesky 2019; Han 2018; King, Pan and Roberts 2017), targeting political dissidents (Gallagher and Miller 2021;

[27]I use stance and sentiment interchangeably in this paper.

Pan and Siegel 2020), hindering collective actions (King, Pan and Roberts 2013), monitoring lower officials (Lorentzen 2014), and rewarding regime supporters (Esberg 2020). This study adds to this understanding by demonstrating that foreign democratic institutions are censored regardless of content sentiment, while discussions of socioeconomic conditions are generally allowed, my study highlights the importance of keeping the public uninformed or misinformed about democratic processes to authoritarian survival.

Moreover, my study addresses the central puzzle of authoritarian institutions critical to understanding the dynamics of authoritarian politics. For a long period, scholars have debated whether the difference between democratic and autocratic institutions is one of degree or kind (Gandhi 2008; Svolik 2012; Truex 2016), particularly as autocratic leaders worldwide strive to claim themselves as truly democratic (Kirsch and Welzel 2019; Shi and Lu 2010; Wu, Weatherall and Huang 2021). My study sheds light on this debate from the angle of authoritarian communication strategy. The strict censorship of foreign democratic institutions found in this study implies that, despite autocrats' efforts to use seemingly democratic institutions, there is deep-seated insecurity among them about the legitimacy of their institutions compared to democratic ones.

3.2 Censorship of Liberal Democracies: Theories

Authoritarian regimes have an intrinsic incentive to uninform or misinform their citizens regarding liberal democracies. This stems from the fear of unfavorable international comparisons, leading the public to admire more advanced democratic societies and political systems (Huang 2015). Such international benchmarking can potentially exacerbate dissatisfaction with domestic governance and ignite public demand for more political power, exemplified by the third wave of democratization following the decline of communism

and the Arab Spring in the early 2010s (Gläßel and Paula 2020; Gleditsch and Ward 2006; Huntington 1991; Steinert-Threlkeld 2017). While conventional wisdom emphasizes the role of censorship in fending off criticism against the government and the autocratic system, I posit that the primary incentive for authoritarian censorship is to minimize public exposure to democratic institutions from online discussions regardless of whether they portray democratic regimes as good or bad.

3.2.1 Blocking State Critique

The conventional approach to understanding how authoritarian regimes counteract the perceived threats from international influences through censorship and propaganda is called *state critique theory*. At its core, state critique theory examines the stance of publications and censors those critical of the authoritarian government or those undermining the legitimacy of the authoritarian state and promotes those in favor of it. In the domestic context, state critique theory posits that censorship primarily targets criticism against the regime, including censoring anti-government speeches (Gueorguiev and Malesky 2019), banning independent media (Lorentzen 2014), targeting influential opinion leaders (Gallagher and Miller 2021), flooding the public forum with pro-regime content (King, Pan and Roberts 2017; Roberts 2018).

In the international context, state critique theory extends to the censorship of international criticism against authoritarian regimes and the negative propaganda of rival regimes. For example, Dukalskis (2021) demonstrates how autocracies like China, North Korea, and Rwanda repress their critical exiles abroad and use combinations of promotional and obstructive tactics to manage their international images. More importantly, authoritarian regimes use international propaganda and censorship to target domestic audiences, systematically depicting liberal democracies as disorderly and dysfunctional (Carter and

Carter 2023), accusing rival democratic regimes of being the “villain" against the people (Mattingly and Yao 2022), strategically blaming the West for their own domestic problems (Rozenas and Stukal 2019), and blocking rival propaganda campaigns infiltrating domestic politics (Golovchenko 2022).

This line of research implies that authoritarian censorship should primarily target positive coverage of foreign liberal democracies, particularly their main rivals, to maintain national pride that could buoy support for the regime (Golovchenko 2022; Greene and Robertson 2022). Such censorship of positive sentiment toward foreign liberal democracies is particularly crucial when authoritarian regimes perform relatively poorly in economic and technological development compared to their democratic rivals (Huang 2015). For instance, toward the end of the Cold War, the governance gap between East and West Germany was widening significantly, and there is historical evidence that the East German government routinely censored news programs from West Germany that showcased better living standards in the West (Gläßel and Paula 2020).

> *Hypothesis 1: Content with a positive stance toward foreign liberal democracies is more likely to be censored than content with a negative stance.*

However, the state critique strategy might not be the most effective censorship tactic. First, after decades of rapid development, the socioeconomic gap between democracies and autocracies has narrowed significantly (Luo and Przeworski 2019). Therefore, positive coverage of the socioeconomic conditions in advanced democracies becomes less threatening to authoritarian governments today compared to the Cold War period. Moreover, the growing tensions between major democratic and autocratic powers, such as the US-China

competition, led to a decline in favorable public opinion toward democratic regimes.[28] As I will demonstrate later in the empirical section, over 70 percent of pre-censorship content regarding liberal democracies on Chinese social media is already negative toward democracies like the US and Japan, whereas less than 20 percent is positive. Therefore, the threat from positive coverage of democracies is relatively low. Finally, censoring political dissidents is costly, likely causing backlash against the censorship activities (Pan and Siegel 2020; Roberts 2020), and more difficult to use automated censorship methods such as keyword blocking, resulting in higher costs in hiring Internet censors (Han 2018; King, Pan and Roberts 2017; Roberts 2018).

3.2.2 Minimizing Exposure to Democracy

In contrast to the state critique theory, I propose a more nuanced understanding of authoritarian censorship strategies regarding liberal democracies. Specifically, I argue that authoritarian regimes censor information about democracies not solely to suppress criticism of the domestic government and admiration of foreign regimes but, more importantly, to keep the public uninformed or misinformed about democratic processes. The essence of such an approach is two-fold. First, citizens in authoritarian regimes have little experience or prior knowledge about democratic institutions and processes, giving authoritarian governments an inherent advantage in manipulating or even monopolizing public understanding of democracy. Second, ordinary citizens in authoritarian regimes often lack both the motivation and the means to acquire knowledge about liberal democracy beyond the available information from official propaganda. As such, keeping the public uniform about democratic institutions and processes is more effective for the autocracy than suppressing critiques.

[28]The fourth wave of the Asian Barometer Survey shows that only 30 percent of Chinese respondents believe American influence on China is positive.

Throughout history, democracy has not been the default form of government, and the vast majority of the regimes were autocratic (Svolik 2012). It was not until the Age of Enlightenment that many concepts of economic and political liberalism, along with the democratic government as we know it today, began to emerge. Democracy did not become the predominant form of government until the fall of communism in the late 20th century. Consequently, ordinary citizens in authoritarian regimes do not necessarily understand how democratic systems function, even as the term “democracy" gains international popularity and becomes synonymous with good governance. This is evident in many of the newly democratized regimes experiencing a period of democratic learning before democracy consolidates (Bratton and Van de Walle 1997; Svolik 2013). Thus, due to such a lack of democratic experiences, authoritarian governments have a natural advantage in manipulating public understanding of democracy.

However, the global waves of democratization did increase pressure on autocracies to adopt nominally democratic institutions to lower the risk of regime breakdown (Gandhi 2008). To capitalize on their advantage of public democratic illiteracy, authoritarian governments spin the meaning of democracy and use nominally democratic institutions to consolidate autocratic rules (Guriev and Treisman 2023; Svolik 2012). In East Asian autocracies, governments often associate the definition of democracy with paternalistic guardianship (Shi and Lu 2010). The Chinese government, in particular, propagates itself as the world’s “greatest democracy” and claims democracy as a “core socialist value” (Wu, Weatherall and Huang 2021). As a consequence, surveys repeatedly find that Chinese citizens regard China as fairly democratic, and in many parts of the world, “support for democracy” actually indicates support for autocracy (Kirsch and Welzel 2019).

Therefore, coverage of democratic institutions and processes abroad becomes particularly threatening to the legitimacy of authoritarian regimes, even when the coverage is

negative. This is because it exposes the public to democratic systems different from the distorted domestic version and educates them about democratic norms and processes. Real democratic institutions and processes are often messy, demonstrated perfectly by no other than contemporary American politics. Examples prevalent in the pre-censorship Chinese social media include the election and impeachment of Donald Trump, the Supreme Court overturning of *Roe v. Wade*, and the conflict between the Federal and Texas governments over the border. Nevertheless, these messy political events underscore crucial democratic values such as free and fair elections, checks and balances, and federalism. While developing democratic values is difficult, once established in society, they might be immune to subpart governance and socioeconomic conditions causing persistent damage to autocrats (Claassen and Magalhães 2022; Svolik 2013). Therefore, authoritarian governments have strong motivations to keep the public uninformed or misinformed about the true meaning of democracy and the details of how democratic institutions and systems function.

A second reason why minimizing public exposure to democracy might be a more effective censorship strategy than suppressing state critique is that ordinary citizens generally have little incentive to understand how foreign democratic institutions function (Chen and Yang 2019). It is well established in the American politics literature that voters have little incentive to learn about political issues, and political knowledge is scant even in democracies with free media (Bartels 1996; Carpini and Keeter 1996). In autocracies, the low salience of political participation leads to even lower incentives for political literacy. For example, less than half of the male respondents in Vietnam can correctly name the National Assembly Chair, while less than 20 percent of female respondents can do so (Schuler 2019), compared to around two-thirds of Americans correctly identifying the parties in control of House and Senate (Borelli and Gracia 2023).

Due to the generally low level of political knowledge, chronic suppression of information related to democratic institutions is less likely to backfire, especially in comparison to the state critique strategy. This is because only censoring state critique would lead citizens to perceive official propaganda as more biased as they can benchmark official information on their own life experiences (Geddes and Zaller 1989). Taking the two arguments together, when alternative sources of information about foreign democracies, i.e., state-sponsored propaganda, are available, most people would not bother paying the extra cost to bypass censorship to access foreign information (Golovchenko 2022; Roberts 2018). Consequently, systematically censoring information about foreign democratic institutions and processes becomes a low-cost and low-salience strategy to effectively uninform or misinform the public regarding democratic values and practices, contributing to authoritarian stability and survival.

A well-known method along the lines of the suppressing exposure strategy is blocking information from abroad altogether using legislative measures, media control, and technologies such as China's "Great Firewall" (Gläßel and Paula 2020; Han 2018; Roberts 2018). Such an isolation strategy was particularly prevalent during the Cold War when international travel and information exchange, especially across the capitalist and the communist blocks, were rare. Yet, the world has significantly globalized since the Cold War. While blanket blockage of foreign information still exists today, particularly in countries like North Korea and Cuba, international knowledge is becoming more accessible to citizens in authoritarian regimes, even with Internet barriers such as China's "Great Firewall" (Huang 2015; Roberts 2018). Moreover, it is theoretically difficult to parse whether censorship strategies like the "Great Firewall" is aimed for fending off state critiques or uninforming the public about the liberal democratic world. Thus, a more fine-grained analysis of the censorship of democracy is needed.

Hypothesis 2: Content related to democratic institutions is more likely to be censored than content related to socioeconomic conditions, regardless of its stance and sentiment.

3.3 Analyzing Chinese Censorship of Democracies

Gauging authoritarian governments' censorship strategies is challenging because of autocracies' secretive nature and the complex institutional structure within the censorship apparatus (Tai and Fu 2020). With only publicly available data, I take an inductive approach by comparing the censored and uncensored content on the Internet and backward inducting government censorship objectives. Although such an approach overlooks the internal dynamic of authoritarian regimes and cannot fully decipher the opaque government intent, it is most effective in revealing the aggregate preference of authoritarian regimes through their censorship behavior (King, Pan and Roberts 2013; Tai and Fu 2020).

The empirical analysis focuses on China, the most sophisticated censorship regime in the world (Roberts 2018). I collected pre-censorship social media articles from *FreeWeChat*, a non-governmental organization monitoring the largest Chinese social media platform, WeChat, with over a billion monthly active users and over 70% penetration rate as of 2023. *FreeWeChat* records WeChat public account articles, similar to Facebook Page posts, and detects censorship by periodically revisiting these recorded articles. Although the selection process for WeChat public accounts is not random but rather by political relevance and levels of activities, studies on social media have repeatedly shown that the vast majority of the content on social media is produced by a fraction of active social media accounts (Tai and Fu 2020). By selectively monitoring active, politically relevant social media accounts, the *FreeWeChat* data captures most major censorship events that shape the average user experience on WeChat.

From all articles recorded by *FreeWeChat* between 2018 and 2022, I use a list of liberal democratic countries' names along with their capitals, major political parties, and recent political leaders to filter out content associated with liberal democracies.[29] This process yielded a corpus of 106,487 articles from 3,074 unique WeChat public accounts. The most frequently mentioned countries are the United States (55.8%), Japan (12.6%), the United Kingdom (11.3%), Germany (7.73%), India (6.82%), France (6.79%), and Korea (5.29%). To alleviate the concern that the corpus is over-concentrated on the United States and ensure that the results are robust across different democratic countries, in the regression analysis, I control for these major countries. Additionally, around half of the articles also mentioned China, which might bias the analysis as domestic topics might be subject to different censorship strategies. To alleviate this concern, in the main analyses, I exclude those mentioning China, while in the appendix, I re-run all analyses using the full dataset. The main findings are generally consistent. Among the articles, 16,908 are recorded to be censored, resulting in an overall censorship rate of 15.88%.

To analyze how the Chinese government censors online content regarding liberal democratic regimes and test the two hypotheses, I first classified all articles related to liberal democracies into different topic categories, including both topics related to democratic institutions and socioeconomic conditions in democratic regimes, and then identified their stance toward liberal democracies. Finally, I compared the censorship rate among different topics and stances to access government censorship strategies.

3.3.1 Identifying Topics

To classify the topic categories, I rely on a semi-supervised keyword-assist topic model (KeyATM) that combines automated content analysis and human-identified concepts

[29] See Appendix A.1 for the full list of keywords.

of substantive interest (Eshima, Imai and Sasaki 2023). Specifically, I identify 12 topics about democratic institutions, 19 about socioeconomic conditions, six about international conflicts, and six about Western history and religions. Appendix A.2 reports the full lists of the keywords. To further minimize the risk of misclassifying content that does not align with the identified topics, the KeyATM model is configured to accommodate an additional 20 topics that are not *a priori* specified. However, no relevant topics related to institutions or socioeconomic status in liberal democracies emerge from these supplementary topics. Table 3.1 lists all specific topic areas. For each article, I assign a specific topic by its highest topic probabilities in the KeyATM model.

Table 3.1: Topics of Online Content Related to Foreign Countries

General Category	Proportion	Specific Topics
Democratic Institutions	28.83%	Checks & Balances, Diaspora Political Participation, Executive (Asia), Executive (Europe), Executive (US), Elections, Federalism & Bureaucracy, Judiciary, Legislature, Media, Economic Liberalism, Political Liberalism
Socioeconomic Conditions	38.99%	COVID-19 Pandemic, Crime & Gun Violence, Cybersecurity, Diaspora Livelihood (US), Diaspora Livelihood (Non-US), Disaster, Economic Performance, Education, Emigrating to Democracies, Energy Policies, Environmental Policies, Healthcare, Housing, Protests (BLM), Stock Market, Technology (AI), Technology (Chips), Technology (EV), Technology (Big Tech Firms)
International Conflicts	12.25%	Conflict in Korean Peninsula, Conflict in Middle East Western Military, Russo-Ukrainian Conflict, Conflict in South Asia, Conflict over Taiwan

Notes: Six additional topics about Western history and religions and 20 additional topics not related to foreign democracies or international affairs are not listed in the table and constitute the remaining 19.92% of the articles.

As a core variable in the hypotheses, identifying topics related to democratic institutions is crucial to the empirical analysis. Among the topics, *Checks & Balances*, *Elections*, *Executive*, *Legislature*, and *Judiciary* are the consensual topics related to democratic institutions according to most canonical works in democracy and democratic institutions (Lijphart 1999; North 1990; Przeworski 2000; Shugart and Carey 1992). Other consensual components of democratic institutions, such as multiple political parties (Tavits 2013) and pluralistic interest groups, are mostly absorbed by the legislature, elections, and executive categories in the corpus. Some canonical works mention federalism as a core element of democratic institutions (Lijphart 1999). In practice, they are intertwined with federal government agencies, such as the Federal Bureau of Investigation, and the 2020 presidential election disputes involving multiple key swing states. Since there is already a topic on elections, I combine federalism and federal government agencies as *Federalism & Bureaucracy*.

Several seminal definitions of democracy also emphasize the importance of civil liberty (Dahl 1989), including freedom of speech, press, and assembly. In my corpus, there are three different types of articles about institutional designs guaranteeing individual freedom. First, some articles discuss the freedom of speech on traditional and social media. Free media is also often regarded as a key pillar of checks and balances. Thus, I include a topic called *Media*. Second, some articles discuss the freedom of assembly by covering protest movements. In the corpus, the vast majority of protest-related articles are about the Black Lives Matter movement. Therefore, it might be biased to categorize them as institutions because racial discrimination is a key aspect of socioeconomic conditions in democracies often propagated in Chinese media. A third type of article discusses the abstract concept of civil liberty, individual rights, and accountable government without extensively covering current events. Thus, I categorize them as *Political Liberalism*.

In addition, economic institutions such as private property rights (North 1990) and independent central bank (Lijphart 1999) are also mentioned as crucial to a well-functioning democratic society. Similar to civil liberty, I categorize the discussions of the abstract concept of private property rights as *Economic Liberalism*. One exception is that a few articles about stock market performance reference the Federal Reserve, a prime example of an independent central bank. Yet, since these articles are predominantly about investment tips, market evaluations, and comments on economic and industrial outcomes, I categorized them as a topic of socioeconomic conditions, namely the *Stock Market*.

Finally, I include an additional topic called *Diaspora Political Participation* to include coverage of ethnic Chinese's involvement in elections, protests, or lawsuits, primarily in the United States, Australia, and Canada. A well-known example is Andrew Yang, who ran for the presidency in 2020 and subsequently for the mayor of New York City in 2021. The reason to parse out ethnic Chinese's political participation is that they potentially have a higher influence on informing the Chinese public about democratic institutions and might be circulated more widely in China.

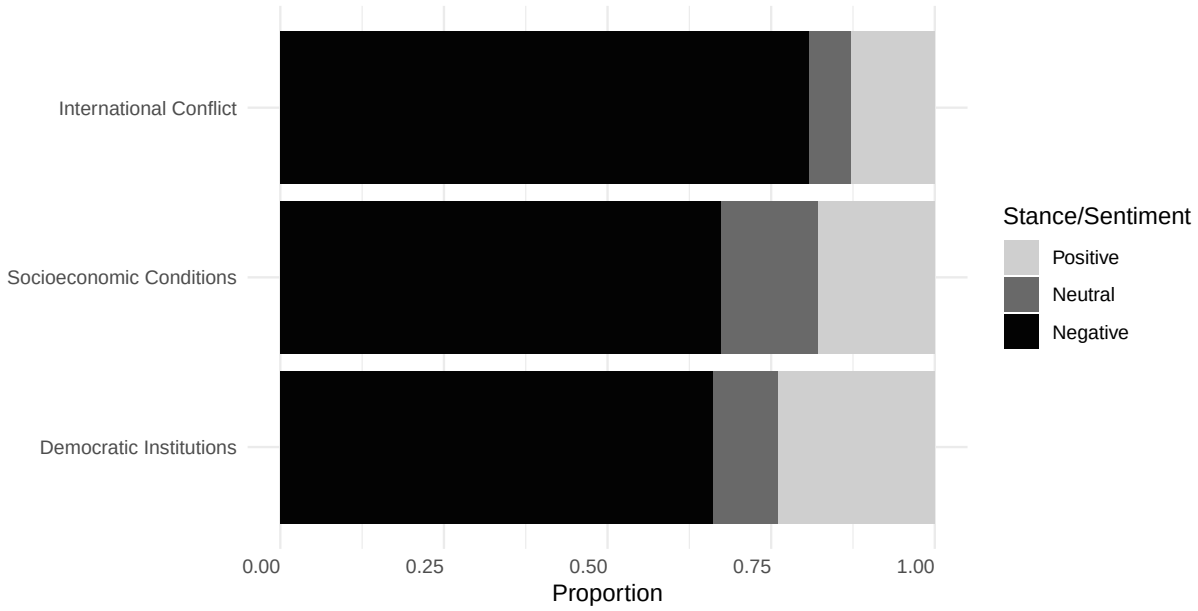
Apart from democratic institutions, as shown in Table 3.1, I identified 19 topics related to socioeconomic conditions and 6 related to international affairs. I only focus on democratic institutions and socioeconomic conditions in the main analysis as they are the most relevant topics to the theoretical arguments and hypotheses. More importantly, democratic institutions and socioeconomic conditions in democracies are less directly related to China's domestic politics and therefore less likely to be affected by other censorship strategies applied to domestic politics. In contrast, content in international conflicts might be related to China's foreign policies and less about liberal democratic regimes.

3.3.2 Identifying Stances and Sentiments

To classify the stances and sentiments of each article, I adopt a supervised approach and use both human annotations and the state-of-the-art deep learning algorithm. I define a positive stance toward liberal democracies as a positive evaluation, positive news, or admiration of the target country or implying that the target country is better than or superior to China. Conversely, I define a negative stance as a negative evaluation, negative news, or resentment of the target democratic country or implying that the target country is worse than or inferior to China. Finally, I define a neutral stance as neither positive nor negative or unclear.

I randomly sampled 3,000 articles from the corpus as the training set and hired two native Chinese research assistants to annotate the training data. The inter-coder reliability is high, with Cohen's $\kappa = 0.95$. I then use this training data to fine-tune the pre-trained Chinese BERT with the Whole Word Masking model, a state-of-the-art deep learning model (Lu, Pan and Xu 2021). Figure 3.1 shows the results of the categorization analysis and the proportion of positive, neutral, and negative articles within each topic category pre-censorship.

Overall, a majority of the pre-censorship content on Chinese social media is negative toward liberal democratic regimes. Across all topics, 71.6% of the articles display some negative evaluations or even resentment of liberal democracies, whereas only 18.2% of the articles are positive and 10.2% neutral. This perhaps reflects the current hostile relationship between China and most democracies, particularly the United States, and further demonstrates the theoretical argument that there is not a pressing need to suppress state critique. Importantly, topics related to democratic institutions and socioeconomic conditions receive almost the exact level of negative coverage. Thus, there is not a clear



Notes: Both censored and uncensored articles are included.

Figure 3.1: Proportion of Pre-Censorship Articles that are Positive, Neutral, and Negative toward Liberal Democratic Regimes by Topic Categories

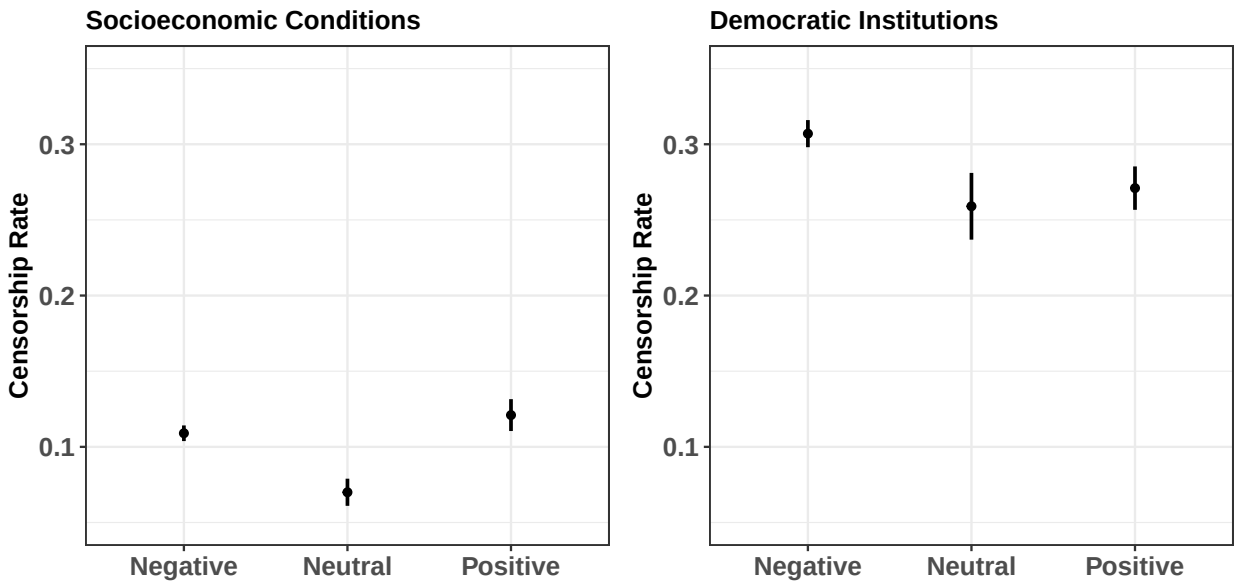
divergence in sentiment between online discourse on democratic institutions and online discourse on socioeconomic conditions.

3.4 Results

The previous section elaborates on the classification of all social media content into different topic areas and stances toward liberal democracies. In this section, I use the difference in censorship rate to gauge Chinese censorship strategies regarding online discourse about liberal democracies. First, I compare the difference in mean between the censorship rates of different subgroups. Then, I run linear regressions to estimate the predictors of censorship incidence.

Figure 3.2 demonstrates the mean censorship rate of content related to socioeconomic conditions (left panel) and democratic institutions (right panel) by their sentiment toward liberal democracies. Consistent with Hypothesis 2, content related to democratic institutions is censored at a significantly higher rate than content related to socioeconomic conditions, with the overall censorship rate for democratic institutions (29.3 percent) three times the overall censorship rate for socioeconomic conditions (10.5 percent). Such a pattern holds regardless of whether the content is positive, negative, or neutral toward liberal democracies. Even when an article has negative sentiment toward democratic institutions, it would still be subject to a higher censorship rate by 18.6 percentage points compared to an article with positive sentiment toward socioeconomic conditions.

In general, content stance plays a minimal role in determining censorship incidences, with the overall difference in means between positive and negative content not statistically significant ($\beta = 0.003$, $p = 0.507$). Although the differences in censorship rates between positive and negative content are statistically significant within socioeconomic conditions categories ($\beta = 0.011$, $p = 0.044$) and democratic institutions categories ($\beta = -0.036$, $p < 0.01$), both differences are much smaller than the gap across democratic institutions and socioeconomic conditions ($\beta = 0.188$, $p < 0.01$). Moreover, the direction of sentiment's



Notes: X-axis indicates whether the content is positive, negative, or neutral toward liberal democracies. Bars indicate the means and the 95 percent confidence intervals of censorship rates. See Table 3.1 for specific topics within democratic institutions and socioeconomic conditions.

Figure 3.2: Censorship Rate of Democratic Institutions and Socioeconomic Conditions by Stance

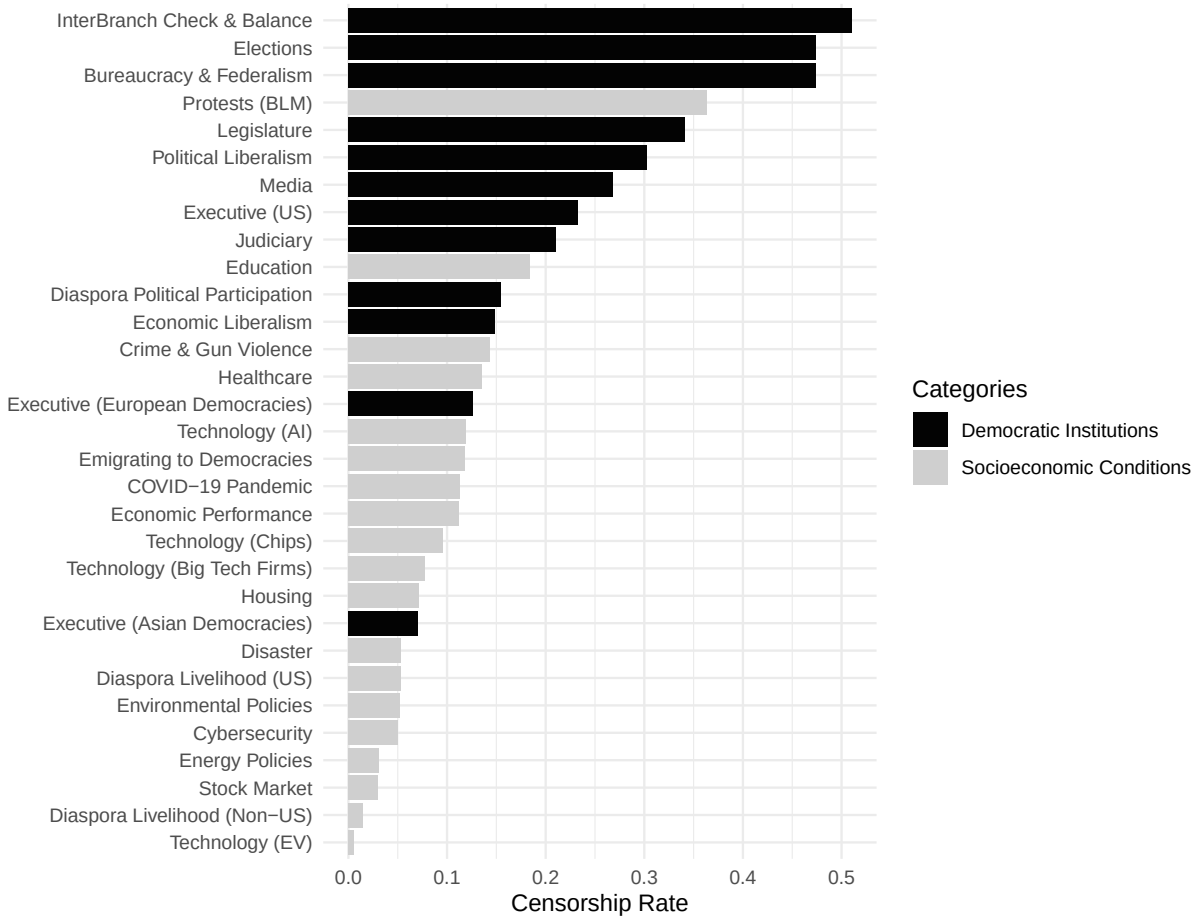
effect is inconsistent across categories, leaving it inconclusive whether positive coverage of democracies is more likely to be censored. In addition, neutral articles are consistently less likely to be censored across categories. This is perhaps because they cover less politically salient events, contributing to lower censorship rates and a lack of clear stance toward democratic regimes. Nevertheless, the differences in censorship probability between neutral content and others are also marginal compared to the differences across topic categories.

To ensure that these main findings are not driven by one single topic and are robust to alternative definitions of democratic institutions, Figures 3.3 and 3.4 demonstrate the censorship rate associated with each specific topic. Consistent with the aggregate results, it is evident that most topics about democratic institutions, including *Elections*,

Checks & Balances, *Legislature*, *Federalism & Bureaucracy*, *Media*, *Judiciary*, and *Political Liberalism* are all subjected to higher levels of censorship. In contrast, the majority of topics about socioeconomic conditions in democracies are tolerated without excessive censorship. Such a lower level of censorship also applies to topics that accentuate the achievement and attractiveness of democratic societies, such as *Education*, *Technology*, *Economic Performance*, and *Emigrating to Democracies*. In fact, had I included *Protests (BLM)* in democratic institutions, the difference in censorship rate between democratic institutions and socioeconomic conditions would become even larger, as *Protests (BLM)* is the fourth highly censored topic category.

The censorship pattern of topics related to the Chinese diaspora communities is also consistent with Hypothesis 2 and the main findings above. Around 15.5 percent of content related to ethnic Chinese participation in democratic elections, political protests, and litigation against discriminatory policies and legislation is censored. This is significantly higher than the censorship rate for content related to the livelihood and socioeconomic conditions of Chinese diasporas in the US (5.3 percent, $\beta = 0.102$, $p < 0.01$), and other democracies (1.5 percent, $\beta = 0.139$, $p < 0.01$).

Another important result in Figure 3.2 is the gap in censorship rates among the three topics related to *Executive*. Articles about the executive branch in the United States are censored at a 23.3 percent rate whereas censorship rates for similar articles about executive branches in Europe and Asia are 12.6 percent and 7 percent, respectively. As China's primary political and strategic rival, it is worth acknowledging that the United States is inevitably over-represented on Chinese social media, and the Chinese regime is indeed applying stricter censorship on content related to the United States. Nevertheless, as demonstrated below, the results are not purely driven by the censorship of the United States, as the



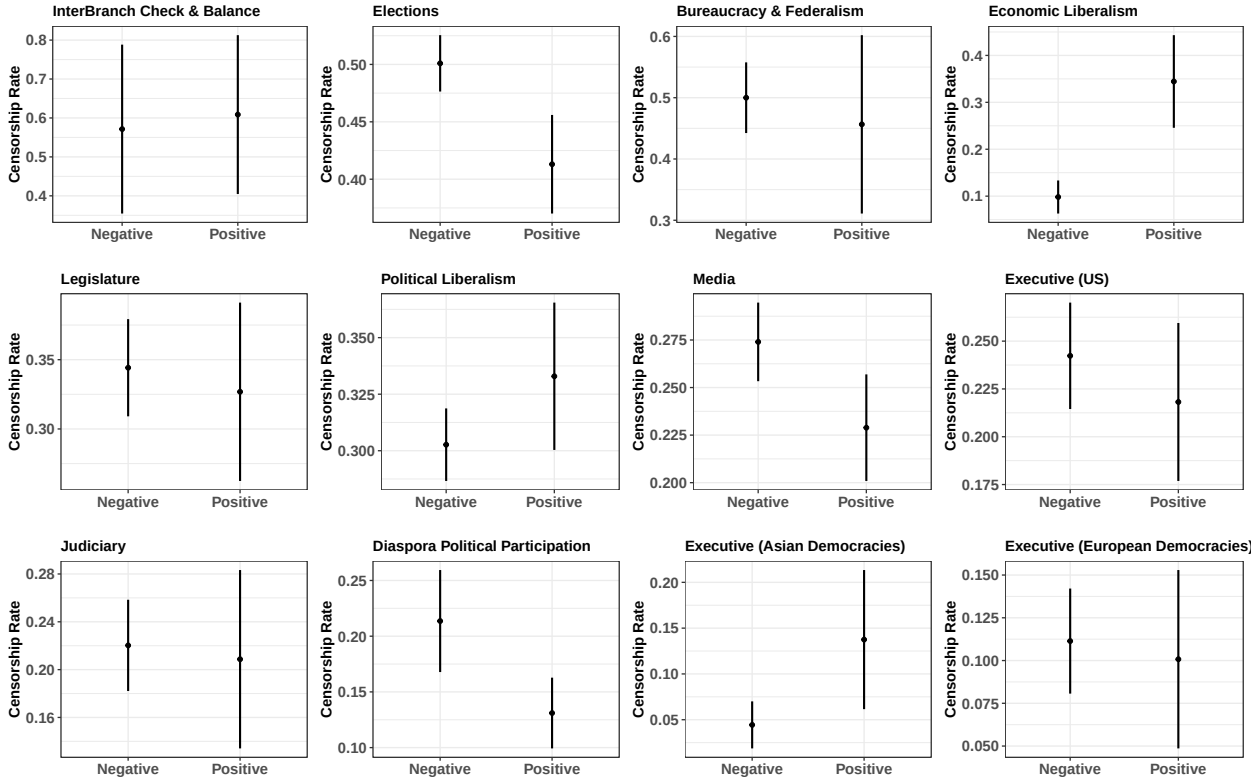
Notes: Bars indicate the overall censorship rate within each topic category regardless of stance.

Figure 3.3: Probability of Censorship by Specific Topic Categories

patterns remain largely unchanged after controlling for all the major countries mentioned in the corpus.

Figure 3.4 further breaks down the censorship rate by specific topic categories within democratic institutions. While the aggregate results shown in Figure 3.2 do not indicate significant discrepancies in censorship patterns between pro- and anti-democracy content, Figure 3.4 indicates substantial heterogeneity among different topics within democratic institutions. Specifically, *Econmic Liberalism* has the highest disparity between positive and negative content (34.4 percent and 9.8 percent, respectively). Articles praising the virtue of

the free market and private property rights are more than three times more likely to be censored than alternative economic views, such as pro-Marxist articles critical of capitalism and the free market. This is partially consistent with Hypothesis 1 and demonstrates that, despite the overall pattern of suppressing strategy, state critique theory is still relevant in the censorship of foreign democracies.



Notes: X-axis indicates whether the content is positive or negative toward liberal democracies. Bars indicate the means and the 95 percent confidence intervals of censorship rates.

Figure 3.4: Probability of Censorship by Specific Topics in Democratic Institutions and Stance

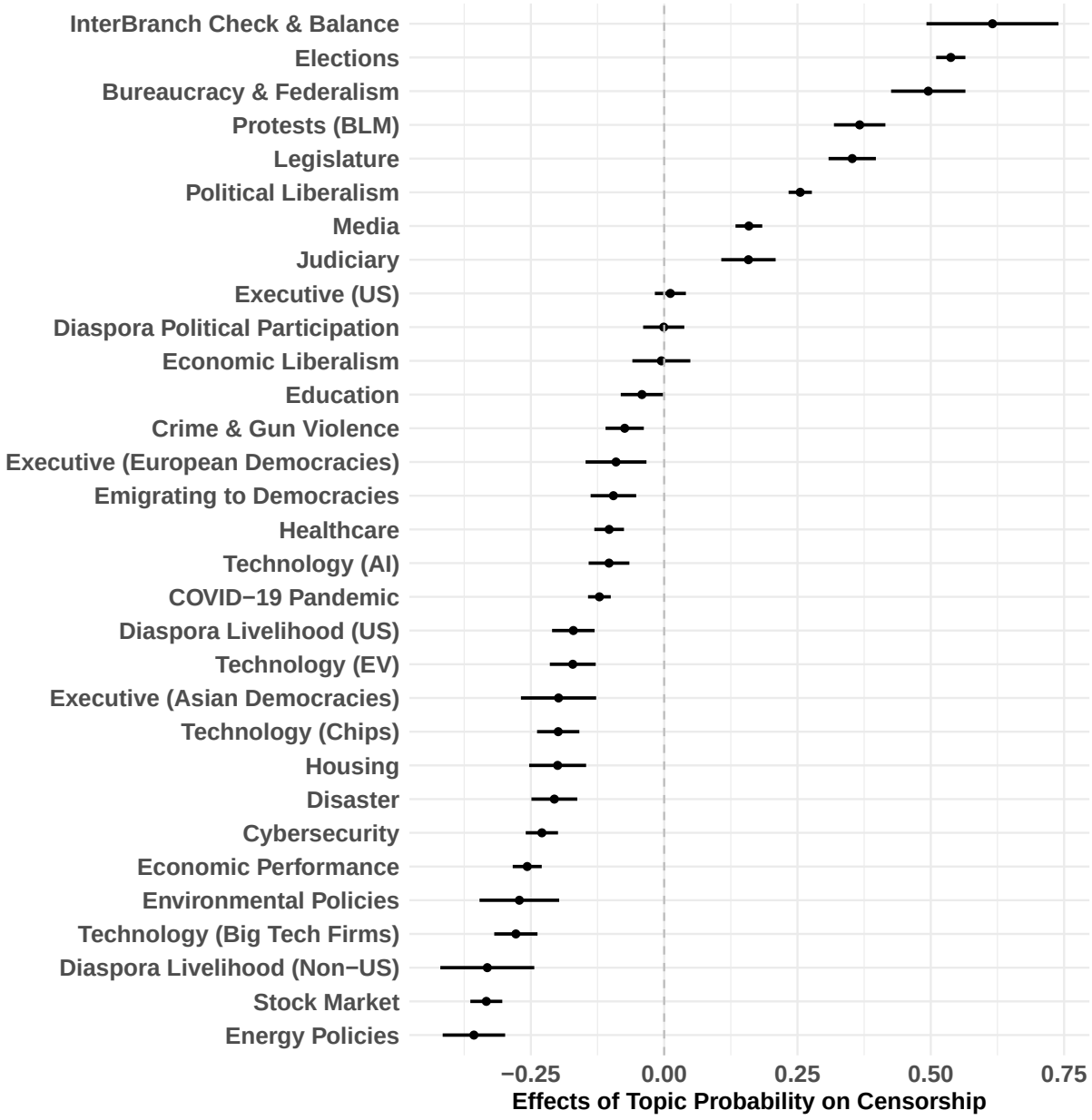
However, one issue with the previous analysis is that I assigned each article to one specific topic according to the highest topic probabilities in the KeyATM result. Yet, in reality, an article could potentially mention multiple topics and the topic with the

highest topic probabilities might not reflect the full article. To address this issue, I use a continuous measure of topic probabilities. For each topic, I then use OLS regressions to illustrate the correlation between that specific topic, stance, and the likelihood of censorship. OLS models simplify the interpretation of results. In the appendix, I use alternative modeling strategies, namely binomial logistic regressions, and show that the results are consistent with the OLS analyses. Additionally, as mentioned above, I control for the major democratic countries mentioned in the articles. Formally:

$$
\begin{aligned}
\text{Censorship}_i \;=\; & \alpha_j + \gamma_j \cdot \text{Topic Probability}_{ij} + \beta_j^N \cdot \text{Stance Neutral}_i + \beta_j^P \cdot \text{Stance Positive}_i \\
& + \delta_j^{US} \cdot \text{Mention US}_i + \delta_j^{UK} \cdot \text{Mention UK}_i + \delta_j^{JP} \cdot \text{Mention JP}_i \\
& + \delta_j^{DE} \cdot \text{Mention DE}_i + \delta_j^{IN} \cdot \text{Mention IN}_i + \delta_j^{FR} \cdot \text{Mention FR}_i \\
& + \delta_j^{KR} \cdot \text{Mention KR}_i + \delta_j^{CA} \cdot \text{Mention CA}_i + \delta_j^{AU} \cdot \text{Mention AU}_i + \epsilon_{ij}
\end{aligned}
$$

where i indicating articles, j indicating topics; and γ_j is the estimates of interest.

Figure 3.5 presents the results of the regression analyses. Overall, the results are consistent with Figure 3.3, demonstrating that the findings are robust to alternative, continuous measures of topics and the target countries. All else equal, an article purely about *Checks & Balances*, or *Elections*, or *Bureaucracy & Federalism* are all 50 percent more likely to be censored compared to an article not mentioning these topics at all, indicating the substantially significant effort by the Chinese government in keep the public uniform about these topics. More importantly, while mentioning the United States does increase the likelihood of censorship, the general pattern of censoring democratic institutions still holds across different liberal democracies.



Notes: Bars indicate the coefficients and the 95 percent confidence intervals of each topic probability. The model is OLS. Control variables include whether the articles mention the United States, the United Kingdom, Japan, India, Germany, France, Canada, Australia, and Korea.

Figure 3.5: Effects of Specific Topic Probability on Censorship Incidence

3.5 Discussion

Despite the censorship pattern demonstrated in the previous section, it is worth emphasizing that propaganda about "democracy" and democratic institutions still exists in authoritarian regimes. In fact, one study analyzes over a million political articles published in the *People's Daily*, the flagship propaganda outlet of the Communist Party of China, and finds that it mentioned "democracy" over 4,000 times per year since it came to power in 1949, equivalent to 77 times per week or 11 times per day (Hu 2020). Therefore, the suppressing exposure to democracy strategy includes both uninforming the public through censorship and misinforming the public through propaganda. While a wealth of literature has investigated the manipulation of public perception and understanding of democracy from the propaganda perspective (Carter and Carter 2023; Dickson 2016; Dukalskis 2021; Hu 2020; Mattingly and Yao 2022; Shi and Lu 2010; Wu, Weatherall and Huang 2021), few have investigated the role of censorship in this misinformation process. Thus, the goal of this study is not to demonstrate that the public knows nothing about democracy, but rather how censorship of democratic institutions contributes to the systematic distortion of democratic norms and values in authoritarian regimes.

Apart from official propaganda, not every public discussion about democratic institutions is strictly prohibited. One of the most prominent examples is the official account of the US Embassy in China, which frequently posts about democratic institutions and processes in the United States. Yet, the US Embassy is rarely censored. Other less influential accounts, such as public intellectuals, sometimes evade censorship while discussing democratic institutions online as well. To understand this discrepancy, it is important to highlight the function of censorship is often not completely erasing unwanted content but rather making it harder to access, a strategy Roberts (2018) called "friction." Complete censorship is often

costly, particularly when the account involves foreign embassies. Therefore, censoring lower-profile discussions about democratic institutions significantly increases the cost for ordinary citizens to obtain this information, screening out most people who do not care enough about it, while avoiding the backlash from high-profile censorship.

3.6 Conclusion

This study provides, to my knowledge, the first empirical analysis of how authoritarian regimes, specifically those like China, censor domestic public discourse on foreign liberal democracies such as the United States and Japan. I propose a novel theory to explain how the Chinese regime censors information about foreign democratic countries. Contrary to the widespread belief that authoritarian regimes like China primarily aim to discredit liberal democracies through their propaganda and censor any positive portrayals of democracies, my findings suggest a more nuanced approach. Drawing on a dataset of over 100,000 pre-censorship articles spanning four years from China's largest social media platform, I demonstrate that discussions about socioeconomic conditions are generally permitted, regardless of whether they convey a favorable or unfavorable perspective on democratic regimes.

In stark contrast, authoritarian regimes exhibit greater vigilance when it comes to the exposure of democratic institutions, even when such exposure carries a negative sentiment and connotation regarding democratic regimes. For the last three decades after the fall of communism in Eastern Europe, authoritarian regimes, notably China, have continued to make strides in economic development and technological advancement, and their dread of prosperous Western societies seems to be dissipating. Future research should further analyze whether the low level of censorship regarding democratic performance persists as

Chinese economic growth slows down and public dissatisfaction with the living standard grows. Nevertheless, despite the persistent official propaganda promoting alternatives to Western liberal democracy, a deep-seated apprehension about democratic systems continues to linger in the minds of autocratic leaders.

Bibliography

Alshabah, Nabli. 2016. "Information Control 2.0: The Cyberspace Administration of China Tames the Internet." *Merics China Monitor* .
URL: *https://merics.org/en/report/information-control-20*

Anderson, Craig A, Akiko Shibuya, Nobuko Ihori, Edward L Swing, Brad J Bushman, Akira Sakamoto, Hannah R Rothstein and Muniba Saleem. 2010. "Violent Video Game Effects on Aggression, Empathy, and Prosocial Behavior in Eastern and Western Countries: A Meta-Analytic Review." *Psychological Bulletin* 136(2):151.

Arendt, Hannah. 1976. *The Origins of Totalitarianism*. Harcourt Brace Jovanovich.

Aronow, Peter M and Allison Carnegie. 2013. "Beyond LATE: Estimation of the Average Treatment Effect with an Instrumental Variable." *Political Analysis* 21(4):492–506.

Bartels, Larry M. 1996. "Uninformed Votes: Information Effects in Presidential Elections." *American Journal of Political Science* pp. 194–230.

Bartholow, Bruce D, Brad J Bushman and Marc A Sestir. 2006. "Chronic Violent Video Game Exposure and Desensitization to Violence: Behavioral and Event-Related Brain Potential Data." *Journal of Experimental Social Psychology* 42(4):532–539.

BBC News. 2020. "China where Fang Fang is From: the Report Culture is Spreading along with the Rise of Nationalism." *BBC News Chinese Version* .
URL: *https://www.bbc.com/zhongwen/simp/chinese-news-52560990*

Borelli, Gabriel and Shanay Gracia. 2023. "What Americans know about their government." *Pew Research Center* .
URL: *https://www.pewresearch.org/short-reads/2023/11/07/what-americans-know-about-their-government/*

Boxell, Levi and Zachary Steinert-Threlkeld. 2021. "Taxing Dissent: The Impact of a Social Media Tax in Uganda." *Available at SSRN* .

Bratton, Michael and Nicolas Van de Walle. 1997. *Democratic Experiments in Africa: Regime Transitions in Comparative Perspective*. Cambridge University Press.

Cairns, Christopher and Allen Carlson. 2016. "Real-world Islands in a Social Media Sea: Nationalism and Censorship on Weibo during the 2012 Diaoyu/Senkaku Crisis." *The China Quarterly* 225:23–49.

Cantoni, Davide, Yuyu Chen, David Y Yang, Noam Yuchtman and Y Jane Zhang. 2017. "Curriculum and Ideology." *Journal of Political Economy* 125(2):338–392.

Carnagey, Nicholas L, Craig A Anderson and Brad J Bushman. 2007. "The Effect of Video Game Violence on Physiological Desensitization to Real-Life Violence." *Journal of Experimental Social Psychology* 43(3):489–496.

Carpini, Michael X Delli and Scott Keeter. 1996. *What Americans Know About Politics and Why It Matters*. Yale University Press.

Carter, Erin Baggott and Brett L Carter. 2023. *Propaganda in Autocracies: Institutions, Information, and the Politics of Belief*. Cambridge University Press.

Chang, Keng-Chi, William R. Hobbs, Margaret E. Roberts and Zachary C. Steinert-Threlkeld. 2022. "COVID-19 Increased Censorship Circumvention and Access to Sensitive Topics in China." *Proceedings of the National Academy of Sciences* 119(4).

Chen, Jidong and Yiqing Xu. 2017. "Why do Authoritarian Regimes Allow Citizens to Voice Opinions Publicly?" *The Journal of Politics* 79(3):792–803.

Chen, Yuyu and David Y Yang. 2019. "The Impact of Media Censorship: 1984 or Brave New World?" *American Economic Review* 109(6):2294–2332.

Cinelli, Carlos and Chad Hazlett. 2020. "Making Sense of Sensitivity: Extending Omitted Variable Bias." *Journal of the Royal Statistical Society Series B-Statistical Methodology* 82(1):39–67.

Claassen, Christopher and Pedro C Magalhães. 2022. "Effective Government and Evaluations of Democracy." *Comparative Political Studies* 55(5):869–894.

CNS. 2020. "2020 Shanghai shi wangluo jubao xuanchuan yue zhengshi qidong [2020 Shanghai Internet Reporting Promotion Month Officially Launched]." *China News Agency* .
URL: *https://www.chinanews.com/sh/2020/09-02/9280670.shtml*

Cook, Sarah. 2019. "Analysis: How the Chinese Communist Party Is Incentivizing Repression." *China Media Bulletin* 133.

Dahl, Robert A. 1989. *Democracy and its Critics*. Yale University Press.

Deng, Rex Weiye. 2023. "Does Negative Propaganda against Foreign Rivals Cultivate Regime-Stabilizing Attitudes? Evidence from China." *Available at SSRN* .
URL: *https://papers.ssrn.com/sol3/papers.cfm?abstract*$_i d = 4478410$

Dickson, Bruce. 2016. *The Dictator's Dilemma: The Chinese Communist Party's Strategy for Survival*. Oxford University Press.

Dikötter, Frank. 2016. *The Cultural Revolution: A People's History, 1962—1976*. Bloomsbury Publishing USA.

Distelhorst, Greg and Yue Hou. 2017. "Constituency Service Under Nondemocratic Rule: Evidence from China." *The Journal of Politics* 79(3):1024–1040.

Dukalskis, Alexander. 2021. *Making the World Safe for Dictatorship*. Oxford University Press.

Efran, Jay S and James E Marcia. 1967. Treatment of Fears by Expectancy Manipulation: An Exploratory Investigation. In *Proceedings of the Annual Convention of the American Psychological Association*. American Psychological Association.

Esberg, Jane. 2020. "Censorship as Reward: Evidence from Pop Culture Censorship in Chile." *American Political Science Review* 114(3):821–836.

Eshima, Shusei, Kosuke Imai and Tomoya Sasaki. 2023. "Keyword-Assisted Topic Models." *American Journal of Political Science* .

Fanti, Kostas A, Eric Vanman, Christopher C Henrich and Marios N Avraamides. 2009. "Desensitization to Media Violence Over a Short Period of Time." *Aggressive Behavior: Official Journal of the International Society for Research on Aggression* 35(2):179–187.

Festinger, Leon. 1957. *A Theory of Cognitive Dissonance*. Vol. 2 Stanford university press.

Freedom House. 2019. "China Media Bulletin: 2019 Internet Freedom Trends, Shutterstock Censorship, Huawei 'Safe Cities' (November 2019).".

Fu, Diana and Greg Distelhorst. 2018. "Grassroots Participation and Repression Under Hu Jintao and Xi Jinping." *The China Journal* 79(1):100–122.

Gallagher, Mary and Blake Miller. 2021. "Who Not What: The Logic of China's Information Control Strategy." *The China Quarterly* 248(1):1011–1036.

Gandhi, Jennifer. 2008. *Political Institutions Under Dictatorship*. Cambridge University Press.

Geddes, Barbara and John Zaller. 1989. "Sources of Popular Support for Authoritarian Regimes." *American Journal of Political Science* 33(2):319–347.

Gläßel, Christian and Katrin Paula. 2020. "Sometimes Less is More: Censorship, News Falsification, and Disapproval in 1989 East Germany." *American Journal of Political Science* 64(3):682–698.

Gleditsch, Kristian Skrede and Michael D Ward. 2006. "Diffusion and the International Context of Democratization." *International Organization* 60(4):911–933.

Goldfried, Marvin R. 1971. "Systematic Desensitization as Training in Self-Control." *Journal of Consulting and Clinical Psychology* 37(2):228.

Golovchenko, Yevgeniy. 2022. "Fighting Propaganda with Censorship: A Study of the Ukrainian Ban on Russian Social Media." *The Journal of Politics* 84(2):639–654.

Greene, Samuel A and Graeme Robertson. 2022. "Affect and Autocracy: Emotions and Attitudes in Russia after Crimea." *Perspectives on Politics* 20(1):38–52.

Gregory, Paul R. 2009. *Terror by Quota*. Yale University Press.

Gueorguiev, Dimitar. 2021. *Retrofitting Leninism: Participation Without Democracy in China*. Oxford University Press.

Gueorguiev, Dimitar D and Edmund J Malesky. 2019. "Consultation and Selective Censorship in China." *The Journal of Politics* 81(4):1539–1545.

Guriev, Sergei and Daniel Treisman. 2015. "How Modern Dictators Survive: An Informational Theory of the New Authoritarianism." *National Bureau of Economic Research* .

Guriev, Sergei and Daniel Treisman. 2023. *Spin Dictators: The Changing Face of Tyranny in the 21st Century*. Princeton University Press.

Han, Rongbin. 2015. "Defending the authoritarian regime online: China's "voluntary fifty-cent army"." *The China Quarterly* 224:1006–1025.

Han, Rongbin. 2018. *Contesting Cyberspace in China: Online Expression and Authoritarian Resilience*. Columbia University Press.

He, Baogang and Mark E Warren. 2011. "Authoritarian Deliberation: The Deliberative Turn in Chinese Political Development." *Perspectives on Politics* 9(2):269–289.

Hobbs, William R and Margaret E Roberts. 2018. "How Sudden Censorship Can Increase Access to Information." *American Political Science Review* 112(3):621–636.

Hu, Yue. 2020. "Refocusing Democracy: the Chinese Government's Framing Strategy in Political Language." *Democratization* 27(2):302–320.

Huang, Haifeng. 2015. "International Knowledge and Domestic Evaluations in a Changing Society: The Case of China." *American Political Science Review* 109(3):613–634.

Huang, Haifeng. 2017. "A War of (Mis)Information: The Political Effects of Rumors and Rumor Rebuttals in an Authoritarian Country." *British Journal of Political Science* 47(2):283–311.

Huang, Haifeng. 2018. "The Pathology of Hard Propaganda." *The Journal of Politics* 80(3):1034–1038.

Huang, Haifeng and Yao-Yuan Yeh. 2019. "Information from Abroad: Foreign Media, Selective Exposure and Political Support in China." *British Journal of Political Science* 49(2):611–636.

Huntington, Samuel P. 1991. "Democracy's Third Wave." *Journal of Democracy* 2:12.

Jiang, Jue. 2021. "The Eyes and Ears of the Authoritarian Regime: Mass Reporting in China." *Journal of Contemporary Asia* 51(5):828–847.

Jost, John T. 2020. *A Theory of System Justification*. Harvard University Press.

King, Gary, Jennifer Pan and Margaret E. Roberts. 2013. "How Censorship in China Allows Government Criticism but Silences Collective Expression." *American Political Science Review* 107(2):326–343.

King, Gary, Jennifer Pan and Margaret E. Roberts. 2014. "Reverse-Engineering Censorship in China: Randomized Experimentation and Participant Observation." *Science* 345(6199):1251722.

King, Gary, Jennifer Pan and Margaret E. Roberts. 2017. "How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, Not Engaged Argument." *American Political Science Review* 111:484–501.

Kirsch, Helen and Christian Welzel. 2019. "Democracy Misunderstood: Authoritarian Notions of Democracy Around the Globe." *Social Forces* 98(1):59–92.

Kostka, Genia. 2019. "China's Social Credit Systems and Public Opinion: Explaining High Levels of Approval." *New Media & Society* 21(7):1565–1593.

Kuran, Timur. 1997. *Private Truths, Public Lies: The Social Consequences of Preference Falsification*. Harvard University Press.

Lijphart, Arend. 1999. *Patterns of Democracy: Government Forms and Performance in Thirty-six Countries*. Yale University Press.

Lorentzen, Peter. 2014. "China's Strategic Censorship." *American Journal of Political Science* 58(2):402–414.

Lu, Yingdan, Jennifer Pan and Yiqing Xu. 2021. "Public Sentiment on Chinese Social Media During the Emergence of COVID-19." *Journal of Quantitative Description: Digital Media* 1:1–47.

Luo, Zhaotian and Adam Przeworski. 2019. "Why are the Fastest Growing Countries Autocracies?" *The Journal of Politics* 81(2):663–669.

Luo, Zhifan and Muyang Li. 2022. "Participatory Censorship: How Online Fandom Community Facilitates Authoritarian Rule." *New Media & Society* .

Lv, Aofei and Ting Luo. 2018. "Asymmetrical Power between Internet Giants and Users in China." *International Journal of Communication* 12:3877–3895.

Manion, Melanie. 2015. *Information for Autocrats: Representation in Chinese Local Congresses*. Cambridge University Press.

Marbach, Moritz and Dominik Hangartner. 2020. "Profiling Compliers and Noncompliers for Instrumental-Variable Analysis." *Political Analysis* 28(3):435–444.

Marcia, James E, Barry M Rubin and Jay S Efran. 1969. "Systematic Desensitization: Expectancy Change or Counterconditioning?" *Journal of Abnormal Psychology* 74(3):382.

Martin, Justin D, Ralph J Martins and Robb Wood. 2016. "Desire for Cultural Preservation as a Predictor of Support for Entertainment Media Censorship in Saudi Arabia, Qatar, and the United Arab Emirates." *International Journal of Communication* 10:23.

Mattingly, Daniel C and Elaine Yao. 2022. "How Soft Propaganda Persuades." *Comparative Political Studies* 55(9):1569–1594.

Miller, Blake. 2018. "Delegated Dictatorship: Examining the State and Market Forces behind Information Control in China" PhD thesis University of Michigan, Ann Arbor.

Mullinix, Kevin J, Thomas J Leeper, James N Druckman and Jeremy Freese. 2015. "The Generalizability of Survey Experiments." *Journal of Experimental Political Science* 2(2):109–138.

Nabi, Zubair. 2014. "Resistance Censorship is Futile." *First Monday* .
URL: *https://firstmonday.org/article/view/5525/4155*

Ng, Jason Q. 2015. "Politics, Rumors, and Ambiguity: Tracking Censorship on WeChat's Public Accounts Platform." *Munk School of Global Affairs* .

Nie, Ke. 2021. "Disperse and Preserve the Perverse: Computing How Hip-Hop Censorship Changed Popular Music Genres in China." *Poetics* p. 101590.

Nimmo, Ben and David Agranovich. 2022. "Meta's Adversarial Threat Report, First Quarter 2022.".

Nisbet, Erik C, Olga Kamenchuk and Aysenur Dal. 2017. "A Psychological Firewall? Risk Perceptions and Public Support for Online Censorship in Russia." *Social Science Quarterly* 98(3):958–975.

North, Douglass C. 1990. *Institutions, Institutional Change and Economic Performance*. Cambridge University Press.

Ong, Lynette H. 2022. *Outsourcing Repression: Everyday State Power in Contemporary China*. Oxford University Press.

Pan, Jennifer and Alexandra A. Siegel. 2020. "How Saudi Crackdowns Fail to Silence Online Dissent." *American Political Science Review* 114(1):109–125.

Pan, Jennifer and Yiqing Xu. 2020. "Gauging Preference Stability and Ideological Constraint under Authoritarian Rule." *21st Century China Center Research Paper, Available at SSRN* .

Pop-Eleches, Grigore and Lucan A Way. 2021. "Censorship and the Impact of Repression on Dissent." *American Journal of Political Science* .

Przeworski, Adam. 2000. *Democracy and Development: Political Institutions and Well-being in the World, 1950-1990*. Cambridge University Press.

Roberts, Margaret E. 2018. *Censored: Distraction and Diversion Inside China's Great Firewall*. Princeton University Press.

Roberts, Margaret E. 2020. "Resilience to Online Censorship." *Annual Review of Political Science* 23(1):401–419.

Rozenas, Arturas and Denis Stukal. 2019. "How Autocrats Manipulate Economic News: Evidence from Russia's State-Controlled Television." *The Journal of Politics* 81(3):982–996.

Schuler, Paul. 2019. "Female Autocrats as Role Models? The Effect of Female Leaders on Political Knowledge and Engagement in Vietnam." *The Journal of Politics* 81(4):1546–1550.

Shadmehr, Mehdi and Dan Bernhardt. 2015. "State Censorship." *American Economic Journal: Microeconomics* 7(2):280–307.

Shi, Tianjian and Jie Lu. 2010. "The Meanings of Democracy: The Shadow of Confucianism." *Journal of Democracy* 21(4):123–130.

Shugart, Matthew Soberg and John M Carey. 1992. *Presidents and Assemblies: Constitutional Design and Electoral Dynamics*. Cambridge University Press.

Sina. 2023. "Zhong yang wang xin ban zhao kai quan guo wang luo ju bao gong zuo hui yi [The Central Cyberspace Administration of China Held a National Online Reporting Work Meeting]." *Sina News* .
URL: *https://finance.sina.com.cn/tech/roll/2023-04-18/doc-imyqumiq8779068.shtml*

Steinert-Threlkeld, Zachary C. 2017. "Spontaneous Collective Action: Peripheral Mobilization during the Arab Spring." *American Political Science Review* 111:379–403.

Stockmann, Daniela. 2013. *Media Commercialization and Authoritarian Rule in China*. Cambridge University Press.

Stromseth, Jonathan R, Edmund J Malesky, Dimitar D Gueorguiev, Lai Hairong and Carl Brinton. 2017. *China's Governance Puzzle: Enabling Transparency and Participation in a Single-Party State*. Cambridge University Press.

Stukal, Denis, Sergey Sanovich, Richard Bonneau and Joshua A Tucker. 2022. "Why Botter: How Pro-Government Bots Fight Opposition in Russia." *American Political Science Review* 116(3):843–857.

Svolik, Milan W. 2012. *The Politics of Authoritarian Rule*. Cambridge University Press.

Svolik, Milan W. 2013. "Learning to Love Democracy: Electoral Accountability and the Success of Democracy." *American Journal of Political Science* 57(3):685–702.

Tai, Yun and King-wa Fu. 2020. "Specificity, Conflict, and Focal Point: A Systematic Investigation into Social Media Censorship in China." *Journal of Communication* 70(6):842–867.

Tavits, Margit. 2013. *Post-Communist Democracies and Party Organization*. Cambridge University Press.

Thurston, Anne F. 1984. "Victims of China's Cultural Revolution: The Invisible Wounds: Part I." *Pacific Affairs* 57(4):599–620.

Truex, Rory. 2016. *Making Autocracy Work: Representation and Responsiveness in Modern China*. Cambridge University Press.

Truex, Rory. 2017. "Consultative Authoritarianism and Its Limits." *Comparative Political Studies* 50(3):329–361.

Tufekci, Zeynep. 2017. *Twitter and Tear Gas*. Yale University Press.

Wang, Dakuo and Gloria Mark. 2015. "Internet Censorship in China: Examining User Awareness and Attitudes." *ACM Transactions on Computer-Human Interaction (TOCHI)* 22(6):1–22.

Wike, Richard and Katie Simmons. 2015. "Global Support for Principle of Free Expression, But Opposition to Some Forms of Speech." *Pew Research Center* 18.

Wu, Hsin-Che, Mark Weatherall and Kai-Ping Huang. 2021. "Propagating 'Democracy' in China? A Two-Way Communication Explanation." *Journal of Contemporary China* 30(130):596–612.

Xinhua. 2023. "2022 nian quanguo shouli wangluo weifa he buliang xinxi jubao 1.72 yi jian [In 2022, 172 million reports of illegal and inappropriate online content nationwide.]." *Xinhua News Agency* .
URL: *http://www.news.cn/politics/2023-02/10/c*$_1$129353484.*htm*

Yang, Guobin. 2009. *The Power of the Internet in China: Citizen Activism Online*. Columbia University Press.

Yang, Jisheng. 2021. *The World Turned Upside Down: A History of the Chinese Cultural Revolution*. Farrar, Straus and Giroux.

Young, Lauren E. 2019. "The Psychology of State Repression: Fear and Dissent Decisions in Zimbabwe." *American Political Science Review* 113(1):140–155.

Zhao, Andy and Zhaodi Chen. 2023. "Let's Report Our Rivals: How Chinese Fandoms Game Content Moderation to Restrain Opposing Voices." *Journal of Quantitative Description: Digital Media* 3.

Zhu, Yuner and King-wa Fu. 2021. "Speaking Up or Staying Silent? Examining the Influences of Censorship and Behavioral Contagion on Opinion (Non-)Expression in China." *New Media & Society* 23(12):3634–3655.

Appendix A

Normalization of Censorship Appendix

A.1 Experiments: Survey Procedure and Descriptive Statistics

A.1.1 Survey Procedure & Pre-Registration

The first survey experiment was conducted in December 2020. The second survey experiment was conducted in December 2022. Both surveys were administered in mainland China by a Shanghai-based Chinese online survey company. The participants were recruited by the survey company and then directed to a US-based website, Qualtrics, where they completed the survey anonymously. Once they completed the survey on Qualtrics, they were redirected back to the survey vendor's platform.

All mainland Chinese citizens above 18 years old are eligible for this study. To make sure that the sample covers a broad range of socioeconomic backgrounds, I put quotas on

gender, region, education, and age. In the end, the quotas successfully yielded samples that reflect the population in terms of gender and region. The age distributions are also pretty close to the demographic considering the fact that younger people under 18 are not eligible for the study. The education quotas alleviate the problem of homogeneous survey participants but fall short of yielding a sample representative of the Internet population.

To further ensure sample quality, I used attention checks to screen the respondents at the beginning of the surveys. About 60% of the respondents passed the attention checks yielding 612 valid responses in Study 1 and 3,314 valid responses in Study 2.

Both survey experiments were pre-registered prior to the implementation of the surveys. The anonymized pre-analysis plan of study 1 can be found here: `https://osf.io/73ej5/files/osfstorage/63f8e804bbc5e502ccf80265`.
The anonymized pre-analysis plan of study 2 can be found here: `https://osf.io/4pg8f/?view_only=51d529726e464cd7a22ff5565fbd2fee`.

A.1.2 Compliance with Ethical Principles of Human Subject Research

Both surveys followed all established principles of human subject research and were approved by the Institutional Review Board (IRB) at the researcher's home institution. Although the IRB exempted both studies from a formal consent form, I still included a consent page and information sheet at the beginning of both surveys. All participants were informed about the purpose, length, and format of the study. All participants need to click "I consent" on the information sheet page before they can proceed. They were allowed to opt out of the study at any point in the survey. Incomplete survey responses were not recorded.

Because the treatment prompt explicitly asked the respondents to **imagine** that they were reading WeChat articles, no deception was used. All articles in both experiments were actual WeChat articles that were censored by WeChat. At the end of both surveys, participants were explicitly told that this was an experimental study and that information in the survey might not be representative of reality.

All respondents were paid by the survey firm at its usual rate for their participation. The survey firm was paid by the researcher of this study. All participants were adults and none of them would be put in a disadvantageous position had they chosen not to participate.

Because both surveys were conducted in China, an authoritarian regime, I paid extra caution to protect respondents' information and responses, so that they would not be negatively affected by the authority due to their participation in this study. I did not ask for personal information that could directly identify participants' identities, such as names, phone numbers, and email addresses. I stored all the responses at Qualtrics via an American institutional account. The study passed the information security review at the researcher's home institution.

A.1.3 Survey Sample

Table A.1: Sociodemographics of the Study Participants and Chinese Internet Users

Sociodemographics		Study 1	Study 2	Chinese Internet Users
Region	East	50.8%	54.5%	46.2%
	Central	19.6%	21.5%	22.1%
	West	21.6%	17.5%	23.3%
	Northeast	7.8%	5.9%	8.4%
Gender	Female	49.7%	49.7%	48.1%
	Male	49.7%	49.8%	51.9%
Education	$\leq$ Junior high	3.6%	3.7%	56.1%
	Senior high	12.6%	16.4%	23.8%
	3-year college	25.2%	36.3%	10.5%
	$\geq$ 4-year college	58.5%	43.3%	9.7%
Age	$\leq$ 19	6.5%	2.6%	23.2%
	20-29	31.4%	27.3%	21.5%
	30-39	45.1%	46.0%	20.8%
	40-49	14.9%	15.9%	17.6%
	$\geq$ 50	2.1%	8.2%	16.9%
Income	$\leq$ 3000	7.8%	6.3%	51.0%
	3000-5000	13.9%	13.8%	21.5%
	5000-8000	38.2%	32.4%	14.3%
	$\geq$ 8000	38.9%	47.0%	13.3%
Occupation	Student	8.3%		26.9%
	Self-employed	13.1%		22.4%
	Corporate employee	34.5%		8.0%
	Corporate management	16.3%		2.9%
	Government employee	2.8%		2.8%
	Professional	12.6%		6.0%
	Manufacturing	4.2%		2.6%
	Service worker	3.6%		4.4%
	Migrant worker	2.0%		4.2%
	Farmer	0.7%		6.3%
	Unemployed & Retired	2.0%		13.5%
Location	Urban	71.9%		71.8%
	Rural	28.1%		28.2%

Note: Data about Chinese Internet users are from *The 45th Statistical Report of Internet Development in China*, issued by China Internet Network Information Center in April 2020.

A.1.4 Balance Table

Table A.2: Balance Table

	Study 1			*Study 2*			*Combined*		
	Control	Treated	*p*	Control	Treated	*p*	Control	Treated	*p*
Female	0.469	0.531	.12	0.484	0.507	.28	0.481	0.513	.10
Age Group	3.863	3.931	.59	4.447	4.342	.16	4.318	4.252	.31
Education	3.407	3.447	.58	3.198	3.196	.96	3.244	3.251	.83
Income	3.221	3.242	.82	3.383	3.338	.34	3.348	3.317	.47
Party Member	0.248	0.274	.46	0.136	0.124	.38	0.161	0.156	.74
Ideology	2.668	2.541	.23	2.363	2.290	.07	2.431	2.345	.03
Pol Interest	4.121	4.085	.74	3.781	3.751	.53	3.856	3.824	.48
Social Media	3.313	3.398	.30	3.523	3.497	.54	3.476	3.476	.99

As shown in Table B.6 and Table A.3, the randomization in general is successful, producing mostly balanced groups. However, there is a slight imbalance in ideology, likely due to chance.

Table A.3: Using Covariates to Predict Treatment

	Treatment		
	Study 1	*Study 2*	*Combined*
Female	0.050	0.023	0.030
	(0.042)	(0.022)	(0.020)
Education	0.018	0.0003	0.005
	(0.028)	(0.017)	(0.014)
Age Group	0.013	−0.009	−0.005
	(0.015)	(0.007)	(0.006)
Income	0.002	−0.008	−0.005
	(0.024)	(0.012)	(0.011)
Ideology	−0.012	−0.023	−0.020**
	(0.016)	(0.012)	(0.009)
Party Member	0.032	−0.015	−0.002
	(0.050)	(0.033)	(0.027)
Political Interest	−0.018	−0.004	−0.005
	(0.018)	(0.010)	(0.009)
Social Media Usage	0.020	−0.005	0.001
	(0.021)	(0.012)	(0.010)
N	593	2,119	2,712

**p $<$ .05; ***p $<$ .01

A.2 Experiments: Additional Analyses

A.2.1 OLS Regressions with Covariates

In this section, I report regression results with all pre-treatment covariates for studies 1 and 2. For the combined sample, because one of the pre-treatment covariates is imbalanced (see Table B.6 and Table A.3), I report regression results for both controlling that imbalanced variable only and all pre-treatment covariates. The results are mostly consistent with the main results reported in the main paper.

Table A.4: Treatment Effects on Support for the Censorship Apparatus

	Support for Censorship Apparatus			
	Study 1	*Study 2*	*Combined*	*Combined*
Treatment	0.213^{***}	0.146^{***}	0.155^{***}	0.162^{***}
	(0.075)	(0.040)	(0.035)	(0.035)
Female	0.121	0.121^{***}		0.128^{***}
	(0.076)	(0.041)		(0.036)
Education	-0.019	-0.034		-0.027
	(0.050)	(0.031)		(0.026)
Age Group	0.016	0.020		0.013
	(0.028)	(0.013)		(0.011)
Income	0.103^{**}	0.106^{***}		0.098^{***}
	(0.044)	(0.022)		(0.019)
Ideology	-0.325^{***}	-0.316^{***}	-0.326^{***}	-0.320^{***}
	(0.029)	(0.022)	(0.017)	(0.017)
Party Member	0.109	-0.151^{**}		-0.054
	(0.090)	(0.060)		(0.049)
Political Interest	-0.038	0.028		0.022
	(0.034)	(0.019)		(0.016)
Social Media Usage	-0.013	-0.006		-0.012
	(0.038)	(0.021)		(0.018)
Constant	4.003^{***}	3.675^{***}	4.172^{***}	3.764^{***}
	(0.247)	(0.159)	(0.048)	(0.134)
N	584	2,088	2,733	2,672
Adjusted R^2	0.203	0.129	0.127	0.143

$^{*}p < .1$; $^{**}p < .05$; $^{***}p < .01$

Table A.5: Treatment Effects on Regime Support: Overall Satisfaction

	Regime Support: Overall Satisfaction			
	Study 1	*Study 2*	*Combined*	*Combined*
Treatment	0.194^{***}	0.086^{**}	0.114^{***}	0.112^{***}
	(0.070)	(0.038)	(0.033)	(0.033)
Female	0.111	0.123^{***}		0.131^{***}
	(0.071)	(0.039)		(0.034)
Education	0.025	-0.029		-0.004
	(0.046)	(0.029)		(0.025)
Age Group	0.016	-0.038^{***}		-0.035^{***}
	(0.026)	(0.012)		(0.011)
Income	0.017	0.086^{***}		0.069^{***}
	(0.040)	(0.021)		(0.018)
Ideology	-0.220^{***}	-0.200^{***}	-0.213^{***}	-0.204^{***}
	(0.027)	(0.020)	(0.016)	(0.016)
Party Member	0.239^{***}	-0.053		0.068
	(0.084)	(0.057)		(0.046)
Political Interest	0.050	0.027		0.046^{***}
	(0.031)	(0.018)		(0.015)
Social Media Usage	-0.036	-0.053^{***}		-0.053^{***}
	(0.035)	(0.020)		(0.017)
Constant	4.043^{***}	4.260^{***}	4.314^{***}	4.168^{***}
	(0.228)	(0.151)	(0.045)	(0.126)
N	592	2,084	2,738	2,676
Adjusted R^2	0.146	0.070	0.066	0.083

$^{*}p < .1$; $^{**}p < .05$; $^{***}p < .01$

Table A.6: Treatment Effects on Regime Support: Central Government

	Regime Support: Central Government			
	Study 1	*Study 2*	*Combined*	*Combined*
Treatment	0.213^{***}	0.066^{*}	0.103^{***}	0.100^{***}
	(0.068)	(0.038)	(0.033)	(0.033)
Female	-0.018	0.156^{***}		0.129^{***}
	(0.070)	(0.039)		(0.034)
Education	0.047	-0.005		0.016
	(0.045)	(0.029)		(0.025)
Age Group	0.007	-0.044^{***}		-0.043^{***}
	(0.025)	(0.012)		(0.011)
Income	0.023	0.060^{***}		0.045^{**}
	(0.040)	(0.021)		(0.018)
Ideology	-0.198^{***}	-0.239^{***}	-0.228^{***}	-0.217^{***}
	(0.027)	(0.020)	(0.016)	(0.016)
Party Member	0.161^{*}	-0.087		0.011
	(0.082)	(0.057)		(0.046)
Political Interest	0.030	0.063^{***}		0.074^{***}
	(0.030)	(0.018)		(0.015)
Social Media Usage	-0.030	0.024		0.0004
	(0.034)	(0.020)		(0.017)
Constant	4.178^{***}	4.005^{***}	4.399^{***}	4.010^{***}
	(0.224)	(0.150)	(0.045)	(0.126)
N	591	2,076	2,730	2,667
Adjusted R^2	0.121	0.094	0.074	0.093

$^{*}p < .1$; $^{**}p < .05$; $^{***}p < .01$

Table A.7: Treatment Effects on Regime Support: Local Government

	Regime Support: Local Government			
	Study 1	*Study 2*	*Combined*	*Combined*
Treatment	0.248^{***}	0.085^{**}	0.120^{***}	0.121^{***}
	(0.077)	(0.040)	(0.035)	(0.035)
Female	0.025	0.101^{**}		0.091^{**}
	(0.078)	(0.041)		(0.036)
Education	0.006	0.003		0.008
	(0.051)	(0.031)		(0.026)
Age Group	0.027	-0.021		-0.023^{**}
	(0.028)	(0.013)		(0.011)
Income	-0.047	0.111^{***}		0.079^{***}
	(0.045)	(0.022)		(0.019)
Ideology	-0.256^{***}	-0.221^{***}	-0.243^{***}	-0.232^{***}
	(0.030)	(0.021)	(0.017)	(0.017)
Party Member	0.196^{**}	-0.105^{*}		-0.002
	(0.093)	(0.060)		(0.049)
Political Interest	0.025	0.033^{*}		0.037^{**}
	(0.034)	(0.019)		(0.016)
Social Media Usage	-0.024	-0.013		-0.018
	(0.038)	(0.021)		(0.018)
Constant	4.313^{***}	3.819^{***}	4.287^{***}	3.950^{***}
	(0.252)	(0.159)	(0.048)	(0.134)
N	580	2,079	2,721	2,659
Adjusted R^2	0.134	0.080	0.075	0.085

$^{*}p < .1$; $^{**}p < .05$; $^{***}p < .01$

Table A.8: Treatment Effects on Willingness to Protest

	Willingness to Protest			
	Study 1	*Study 2*	*Combined*	*Combined*
Treatment	-0.293^{***}	-0.082	-0.130^{***}	-0.124^{***}
	(0.103)	(0.051)	(0.046)	(0.046)
Female	-0.124	-0.099^{*}		-0.107^{**}
	(0.105)	(0.053)		(0.047)
Education	0.012	0.055		0.015
	(0.069)	(0.040)		(0.034)
Age Group	0.067^{*}	-0.014		-0.002
	(0.038)	(0.016)		(0.015)
Income	-0.033	0.093^{***}		0.095^{***}
	(0.060)	(0.028)		(0.025)
Ideology	-0.059	-0.130^{***}	-0.136^{***}	-0.129^{***}
	(0.040)	(0.027)	(0.022)	(0.022)
Party Member	0.409^{***}	-0.128^{*}		-0.056
	(0.124)	(0.077)		(0.064)
Political Interest	-0.099^{**}	-0.012		-0.046^{**}
	(0.046)	(0.024)		(0.021)
Social Media Usage	-0.035	-0.077^{***}		-0.054^{**}
	(0.051)	(0.027)		(0.024)
Constant	3.325^{***}	3.414^{***}	3.430^{***}	3.470^{***}
	(0.338)	(0.203)	(0.062)	(0.175)
N	590	2,080	2,734	2,670
Adjusted R^2	0.030	0.034	0.016	0.028

$^{*}p < .1$; $^{**}p < .05$; $^{***}p < .01$

A.2.2 Heterogeneous Treatment Effect

Figure A.1 shows the heterogenous treatment effects among different demographic subgroups. As shown in the figure, treatment effects are weaker among respondents with lower education. This might indicate that (1) lower educated respondents are less able to pick up the treatment or (2) they are less susceptible to normalization. In the meantime, the confidence intervals of lower educated respondents are wider, suggesting that the weaker treatment effect might be due to insufficient sample size.

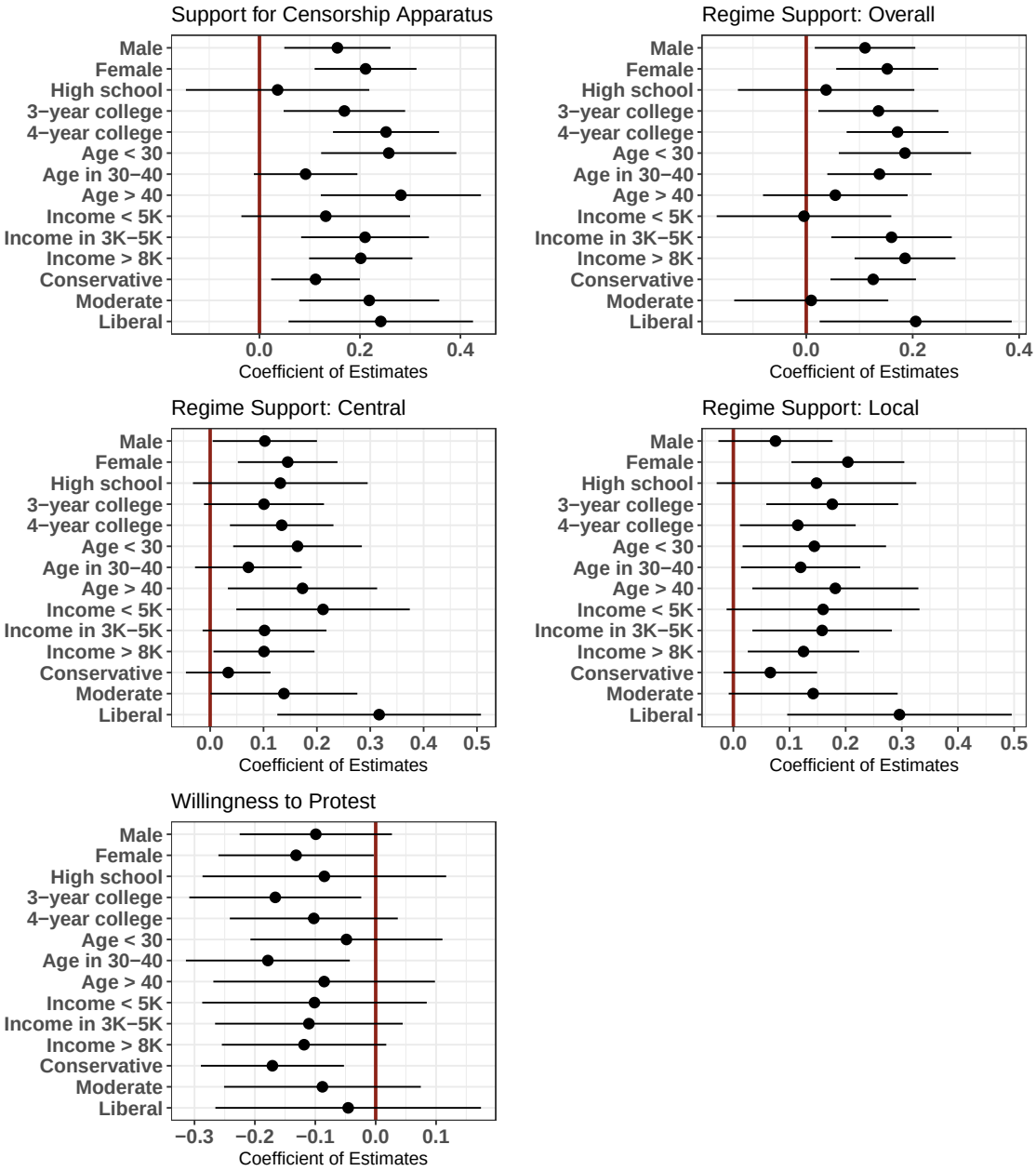


Figure A.1: Heterogeneous Treatment Effects on Outcome Variables (Two Studies Combined)

A.2.3 Multiple Hypotheses Testing Correction

I used the Benjamini-Hochberg (BH) *p*-value correction method to account for multiple comparisons. As shown in Table A.9, all existing significant results survive the BH correction and are still significant at the conventional level.

Table A.9: Multiple Hypotheses Testing Correction (Benjamini-Hochberg)

	Treatment Effects		
Dependent Variable	*Study 1*	*Study 2*	*Combined*
Support for Censorship Apparatus	0.264	0.163	0.185
p-value	[0.00144]	[0.00010]	[0.00000]
adjusted *p*-value	[0.00270]	[0.00043]	[0.00001]
Overall Satisfaction of China	0.229	0.106	0.133
p-value	[0.00183]	[0.00604]	[0.00011]
adjusted *p*-value	[0.00306]	[0.00906]	[0.00043]
Assessment of Central Government	0.236	0.092	0.124
p-value	[0.00097]	[0.01821]	[0.00031]
adjusted *p*-value	[0.00208]	[0.01951]	[0.00092]
Assessment of Local Government	0.288	0.100	0.141
p-value	[0.00037]	[0.01457]	[0.00012]
adjusted *p*-value	[0.00092]	[0.01681]	[0.00043]
Willingness to Protest	−0.268	−0.076	−0.118
p-value	[0.00852]	[0.13341]	[0.01021]
adjusted *p*-value	[0.01162]	[0.13341]	[0.01276]

A.2.4 Implicit Support for Censorship

In addition to the additional analyses in Section 5.5 of the main paper, to further alleviate the concerns about preference falsification, I also use a list experiment to measure implicit support for censorship in Study 2. The list experiment uses the exact same wording as the censorship support question. Figure A.2 the results for implicit and explicit support for the censorship apparatus in Study 2, where explicit support is the proportion of respondents who chose somewhat or strongly support, in the explicit question. If we only examine the point estimates, in the control group, around 46% of the respondents exhibit implicit support for censorship, whereas in the treatment group, this figure rises to 51%. Comparing these results to the explicit support for censorship at 53% for the control group and 58% for the treatment group, two key observations can be made. First, the overall level of preference falsification is at most 7 percentage points. This suggests that, on the whole, preference falsification may not be a pervasive issue. Second, we still detect a similar increase in implicit support when moving from the control group to the treatment group, indicating that the level of preference falsification is not higher in the treatment group. However, the biggest problem with the current list experiment is that their estimates are imprecise. Specifically, the 95% confidence interval for implicit support in the control group is [37%, 57%], and [41%, 62%] for the treatment group. These intervals encompass over 20 percentage points, making it challenging to find any statistically significant treatment effects using implicit measures.

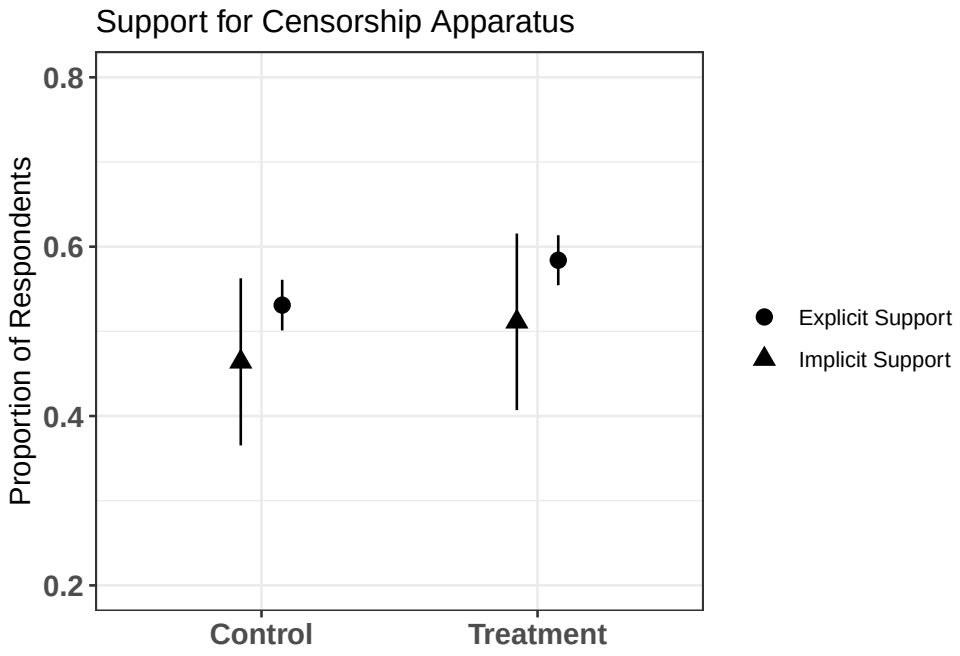


Figure A.2: Implicit Support for Censorship

A.3 Experiments: Experiment Articles

As explained in the main paper, in both experiments, to expose participants to censorship, I asked respondents to read ten snippets of WeChat articles, presented one at a time with only the title and the first few lines. The snippets are screenshots of real articles censored by WeChat. They only include the first couple of lines and do not reveal the full content of the articles. Table A.10 and Table A.11 report the titles of articles used in both studies. For the first experiment, I selected snippets from the WeChatScope dataset used in the observational study. For the second experiment, I selected snippets from another website recording Chinese censorship. The selection process was systematic.

Table A.10: Treatment Articles for Study 1 (Order Randomized)

#	Political Content	Control Group	Treatment Group	Title
1	No		Censorship Label	The banks are crying, one trick to help you earn 23 times higher profit by demand deposit.
2	No		Censorship Label	How much do men care about your face?
3	No		Censorship Label	The King of traditional medicine Sun Simiao lived 142 years. Before he died, he told his pupils: Be sure to destroy this prescription!
4	No			Ten questions about Huawei's former employees being sued for extortion.
5	No			Say goodbye to the stressful status quo. How can we relax under the pressure of work?
6	No			Thaksin and Yingluck returned to Meizhou to worship their ancestors.
7	Yes	Censorship Label	Censorship Label	Just now, the Pingshan Jasic labor strike has won an initial victory!
8	Yes	Censorship Label	Censorship Label	Jiangxi's "Funeral Reform" must not smash the coffin and hurt people's hearts
9	Yes	Censorship Label	Censorship Label	President Hu and Premier Wen are getting old. What happened in their decade?
10	Yes			After the tax reform, has your income decreased?

Table A.11: Treatment Articles for Study 2 (Order Randomized)

#	Political Content	Control Group	Treatment Group	Title
1	No		Censorship Label	Please! Shut up! Don't Like Them Anymore!
2	No		Censorship Label	If you feel pain here, maybe problems with the meridian. Try this herb!
3	No		Censorship Label	Full-time housewife for 20 years and only get 50K for divorce. Brutal truth about marriage.
4	No			Please don't over-interpret Lu Daosen's suicide.
5	No			Wu Zhihong: Be careful of people with too much positive energy.
6	No			Why "Akita beauties" are not happy?
7	Yes	Censorship Label	Censorship Label	24 hours later, is the Xuzhou chained women in black doing okay?
8	Yes	Censorship Label	Censorship Label	Zhang Weiying: Democracy is a commitment.
9	Yes	Censorship Label	Censorship Label	White elephant projects are not government accomplishment.
10	Yes			Chen Jizhi said sorry after convicted, the same as Hu Xijin.

A.4 Text Analysis: Categorization of Censored Articles

A.4.1 Categories and Coding Process

In total, I keep track of nine different topic categories. In addition to three highly political categories: (1) collective action, (2) government criticism, and (3) other government-related articles, I also include six moderately political and non-political categories: (1) business, (2) foreign events, (3) entertainment, (4) advertisement, (5) cultures, and (6) others. The practice of distinguishing non-political content from political content is consistent with recent research on authoritarian censorship (Esberg 2020).

The categorization process and coding rubrics mainly follow Miller (2018), because Miller (2018) provides the most detailed, reliable, and up-to-date categorization of censored content in China. In particular, the definition of collective action, business, and entertainment is the same as Miller (2018). The definition of government-related content combines the definitions of seven different categories in Miller (2018): *government*, *corruption*, *sensitive anniversary*, *recurring political event*, *regular political event*, *nationalism*, and *HK/Macau/Taiwan*. By using a broader definition of government-related content, I aim to establish the upper bound on the proportion of political content and avoid underestimating government-related content. The difference between government criticism and other government-related articles also follows the definition of government criticism in Miller (2018). Any government-related content that does not meet the definition of government criticism is categorized as other government-related articles.

The last four categories were created by myself due to the incompleteness of Miller's coding rules to my data. They are all self-explanatory. Importantly, none of these categories include politically salient events or issues. The coding rubrics for non-political categories explicitly exclude content related to the Chinese government. For example, the business category excludes government economic policies, state-owned enterprises, and any mention of government institutions; the foreign events category requires the article to have no direct reference to China. The last category is the residual category which includes all articles that do not fit into the definitions of the other groups.

One important difference from Miller (2018) is that the nine categories are mutually exclusive. A similar strategy is employed by King, Pan, and Roberts (2013). Having mutually exclusive categories simplifies the categorization process as well as the interpretation of the results. In practice, the nine categories are coded sequentially with political categories coded first. Specifically, an article will first be considered if it belongs to the collective action category. If yes, then the categorization process ends. If not, the article will then be considered if it belongs to the government criticism category and so on. If an article does not fit into the definitions of the first eight categories, it will be put into the last residual category. The coding process ensures that the analysis will not underestimate collective action and government criticism.

A.4.2 Inter-Coder Reliability

Two coders coded the 12,500 articles and posts (2,500 from WeChatScope and 5,000 from FreeWeChat and WeiboScope respectively) in the training set independently. To code the training set, they both analyze the titles, the authors, and the content of the articles according to the coding rubric. Both coders are native Chinese graduate students in

political science. Table A.12 shows that their results are generally consistent in terms of the proportion of each topic category. The greatest disagreement between the two coders is whether an article belongs to Government Criticism (CRI) or other government-related articles (GOV), which is not the main focus of this paper.

Table A.12 shows the details of the two coders' coding. The accuracy rate between the two coders is 82.5% when considering specific topic categories. When identifying whether an article is political or non-political, the two coders agree on 92.97% of the cases. The macro F1 is 0.82 and the Cohen's κ between the two coders is 0.80, higher than the commonly applied criteria of 0.70 for inter-coder reliability tests. In cases where the two coders disagreed, the author acted as an arbitrator to settle the dispute.

Table A.12: Inter-Coder Reliability

	ADS	BET	COL	CRI	ESX	FOR	GOV	LCT	OTH	Macro
Precision	0.92	0.87	0.88	0.94	0.89	0.97	0.68	0.84	0.50	0.83
Recall	0.86	0.90	0.91	0.71	0.89	0.78	0.87	0.79	0.76	0.83
F1	0.89	0.88	0.90	0.81	0.89	0.86	0.77	0.81	0.60	0.82

Note: ADS: Advertisement. BET: Business. COL: Collective Action. CRI: Government Criticism. ESX: Entertainment. FOR: Foreign Events. GOV: Government (Others). LCT: Cultures. OTH: Others.

A.4.3 Content within Each Topic Categories

To better understand what kinds of content are being censored in each specific topic category, I run simple Structural Topic Models (STM) within each of the topic categories. I then manually identified the 10 most common topics among the STM results, along with their associated keywords.

Collective Action

- Topic 1: Hong Kong protest movement in 2019
  - Keywords: HK, violent protesters, riot, protest, police, terrorists, looting
- Topic 2: Other protests in Hong Kong
  - Keywords: HK, Pan Democrats, July 1st, gather, independence, LegCo
- Topic 3: Labor strikes
  - Keywords: labor union, workers, factory, Shenzhen, employee, stop production
- Topic 4: Historical revolutions in China
  - Keywords: revolution, Opium War, Boxer Rebellion, Red Guard, the West
- Topic 5: Uyghur unrest
  - Keywords: Xinjiang, Uyghur, sovereignty, riot, public security
- Topic 6: Weiquan movement and petitioning
  - Keywords: rightful resistance, petitions

- Topic 7: COVID-related collective actions
  - Keywords: COVID, quarantine, testing, Fangcang hospital
- Topic 8: Picking quarrels and provoking trouble
  - Keywords: bully, violence, beer bottle, while shirt, police, incident
- Topic 9: Foreign involvement in collective actions
  - Keywords: spy, CIA, United States, US Congress, Trump
- Topic 10: Bank-related collective actions
  - Keywords: deposit, bank, gather, chanting, migrant workers

Government Criticsm

- Topic 1: COVID-19 pandemic (initial outbreak)
  - Keywords: Li Wenliang, lockdown, Wuhan, Fangcang hospital, Fang Fang
- Topic 2: COVID-19 pandemic (criticism of zero-COVID policy)
  - Keywords: vaccine, zero-COVID, Omicron, Pfizer, case, positive case
- Topic 3: Corruption
  - Keywords: bribery, violation of discipline, take bribes, discipline inspection
- Topic 4: Criticism of the one-child policy
  - Keywords: one-child policy, aging population, birth rate, population growth
- Topic 5: Criticism of the Xiong'An (planned new capital) policy

  - Keywords: Xiong'An, Xiong County, new district, demolition, flooding, villagers

- Topic 6: Criticism of foreign policies
  - Keywords: wolf warriors, little pinky, Hu Xijin,
- Topic 7: Feminism
  - Keywords: women's rights, patriarchy
- Topic 8: Democratic values
  - Keywords: freedom, human rights, liberalism, property rights
- Topic 9: Criticism of the police and censorship
  - Keywords: questioning, the police, netizens, demand, media, investigation, misinformation, censorship, rumor
- Topic 10: Criticism of economic policies
  - Keywords: labor force, employment, pension, fiscal policy, manufacturing

Other Government-Related

- Topic 1: Party leaders
  - Keywords: Xi Jinping, General Secretary, Hu Jintao, Jiang Zemin, Zhu Rongji
- Topic 2: Communist Party
  - Keywords: Party organization, secretary, comrade, appoint, decide, economic development, work, meetings, reform

- Topic 3: Court
  - Keywords: Supreme People's Court, Appeal, Defendant, Laywer, Imprisonment
- Topic 4: Taiwan
  - Keywords: Taiwan, brainwash, defame, anti-China, nation, reunification
- Topic 5: Xinjiang
  - Keywords: Xinjiang, autonomous region, Hetian, safeguard, Terrorists
- Topic 6: Ideology
  - Keywords: Marxism, Mao Zedong, Dong Xiaoping, socialism, capitalism
- Topic 7: History of the Communist Party
  - Keywords: Yan'an, Chairman Mao, Lin Biao, Kuomintang, WWII, history
- Topic 8: Military
  - Keywords: PLA, tank, helicopter, navy, fighter jets, air force
- Topic 9: COVID-19 Pandemic
  - Keywords: vaccine, virus, WHO, immune, mutation, case
- Topic 10: US-China Trade War
  - Keywords: Huawei, trade war, Trump, sanction, chips, RMB

Business

- Topic 1: Investment Tips (Stock Market)

  - Keywords: Index Fund, Stock price, Bond, IPO, Long, Short, Buy-in, Sell-out

- Topic 2: Investment Tips (Real Estate)
  - Keywords: Housing price, Second-hand house, Buying house, School district
- Topic 3: Investment Tips (Crypto Currency)
  - Keywords: Cryto, Dogecoin, Encrypted, Block-chain, Mining, Elon Mask
- Topic 4: Sector Analysis (Platform Companies)
  - Keywords: Meituan, Rider, Kuaishou, JD.com, Douyin, Pinduoduo, Tencent
- Topic 5: Sector Analysis (Food Industry)
  - Keywords: Ruixin Coffee, Starbucks, Nestle, Mooncakes, Brand, Ice cream, Yili
- Topic 6: Sector Analysis (Alcohol Industry)
  - Keywords: Maotai, Alcohol, Baijiu Liquor, Fermentation, Baijiu Aroma, Beer
- Topic 7: Sector Analysis (Electric Cars)
  - Keywords: Tesla, NIO, XPeng, New energy, Electric cars, Self-driving
- Topic 8: Sector Analysis (Education Industry)
  - Keywords: New Oriental, Yu Minhong, Education, Training, Extracurricular
- Topic 9: The Effect of COVID on Investment
  - Keywords: Vaccine, Virus, Coronavirus, Global, Economy
- Topic 10: Investment Tips (Trust & Equity)
  - Keywords: Trust, Billion Yuan, Shareholder, Bank

Foreign Events

- Topic 1: Domestic Politics of the United States
  - Keywords: Democrats, Republicans, midterm election, racism, BLM, conservative
- Topic 2: Domestic Politics of European Countries
  - Keywords: Germany, Euro, Italy, France, Spain, European Union
- Topic 3: Domestic Politics of Russia & Russo-Ukrainian Conflict
  - Keywords: Russia, Putin, Crimea, Ukraine, Donbas, Kyiv
- Topic 4: Overseas Chinese Community Information (US)
  - Keywords: New York, Flushing, Brooklyn, Queens, Chinatown, Los Angeles, Southern California, house rental, restaurants
- Topic 5: Israel-Palestinian Conflict
  - Keywords: Israel, Hamas, Palestine, Syria, Gaza
- Topic 6: Other Events in the Middle East
  - Keywords: Iran, Iraq, Shia, Sunni, Ali Khamenei
- Topic 7: COVID-19 Pandemic in Foreign Countries
  - Keywords: mask, vaccine, infection, CDC, Delta, Omicron
- Topic 8: Overseas Chinese Community Information (Europe)
  - Keywords: Madrid, Catalonia, the Spanish-Chinese community

- Topic 9: Domestic Politics of India
  - Keywords: Modi, India, New Delhi, Nepal, Indian government
- Topic 10: Domestic Politics of Korea & Japan
  - Keywords: Yoshihide Suga, Shinzo Abe, Moon Jae-in, LDP

Entertainment

- Topic 1: Discussion of Movies
  - Keywords: Douban, movies, actors, actress, critics, movie festival, box office, Ashes of Time, Academy Awards
- Topic 2: Discussion of Entertainment Shows
  - Keywords: reality shows, entertainment shows, Xiao Zhan, fans, stars, Sina Weibo, Guo Degang, New Year's Gala
- Topic 3: Tabloid Gossips
  - Keywords: gossip, paparazzi, fans, Zhao Wei, Guo Jingming, Xiao Yaxuan, Lin Zhixuan, Nicholas Tse
- Topic 4: Discussion of Classical Novels
  - Keywords: Dream of the Red Chamber, The Legend of the Condor Heroes, Wolf Totem, The Three-Body Problem
- Topic 5: Discussion of Beauty Standard
  - Keywords: whitening, loss of weight, skin, model

- Topic 6: Discussion of TV Series
  - Keywords: House of Cards, Prison Break, Breaking Bad, Spartacus
- Topic 7: Discussion of Documentaries
  - Keywords: documentaries, BBC, world, Renaissance, WWII, history
- Topic 8: Discussion of Music
  - Keywords: popular music, Wang Mingquan, Tie Xue Dan Xin
- Topic 9: Discussion of Relationships
  - Keywords: husband, wife, boyfriend, girlfriend, love, life, marriage
- Topic 10: Discussion of Talk Shows
  - Keywords: Liang Wendao, Gao Xiaosong, Behind the Headlines with Wen Tao

Advertisement

- Topic 1: Product Promotion: Food
  - Keywords: taste, crawfish, meat, chicken feet, eel, strawberry, corn, fruit
- Topic 2: Product Promotion: Courses & Training
  - Keywords: textbook, vocabulary, multi-media, third grade, training, register
- Topic 3: Job Ads
  - Keywords: hiring, written examination, salary, earnings, position, qualifications
- Topic 4: Product Promotion: Household Product

  - Keywords: teapot, table, mask, sanitizer, toothpaste

- Topic 5: Product Promotion: Beauty Products
  - Keywords: sunscreen, lipstick, moisturizer, face mask, skin
- Topic 6: Product Promotion: Clothing
  - Keywords: material, pants, underwear, T-shirt, lightweight
- Topic 7: Product Promotion: Beverage
  - Keywords: tea, white tea, black tea, Pu'er tea
- Topic 8: Product Promotion: Medical Service
  - Keywords: eye hospital, gout, diabetes, nearsighted, orthopedics
- Topic 9: Product Promotion: Prescription Drugs
  - Keywords: ointment, bacteria, analgesics
- Topic 10: Product Promotion: Herbal Medicine
  - Keywords: Chinese medicine, cordyceps, goji berry, moisten the lungs, drink

Cultures

- Topic 1: Stories of Traditional Chinese Medicine
  - Keywords: master, apprentices, traditional medicine, secret recipe
- Topic 2: Stories of Chinese Poets
  - Keywords: Su Shi, Su Dongpo, Wang Anshi, Li Shangyin, Su Zhe, Tang Dynasty

- Topic 3: Feng Shui Stories
  - Keywords: Feng Shui, secret, fortune, Chinese Zodiac, culture
- Topic 4: Stories of Fictional Figures in Chinese Literature
  - Keywords: Lin Daiyu, Jia Baoyu, Xue Baochai, Jin Ping Mei,
- Topic 5: Stories of Chinese Emperors
  - Keywords: Yongzheng, Kangxi, Qianlong, Liu Bang, Xiang Yu
- Topic 6: I Ching Stories
  - Keywords: I Ching, hexagram, wisdom, culture
- Topic 7: Bible Stories
  - Keywords: God, Jesus, Jehovah, Christ, Gospel
- Topic 8: Buddhist Stories
  - Keywords: Heart Sutra, Sarira, bodhi, incantation
- Topic 9: Taoist Stories
  - Keywords: Taoism, Lao Zi, Zhuang Zi
- Topic 10: Local Architecture
  - Keywords: architecture, photography, art, urban, history

A.5 Text Analysis: Models & Robustness

A.5.1 Model Selection

To select the best classification model, I used the training data to test nine different machine-learning models. As shown in Table A.13, the fine-tuned pre-trained Chinese BERT with the Whole Word Masking model is by far the best-performing model evaluated by out-sample five-fold cross-validation macro F1 score.

Table A.13: Macro F1 Scores for Five-fold Cross-Validation

Model	Macro F1 Score
Fine-tuned Pre-Train Chinese BERT	0.7025
Logistic Regression (Ridge)	0.4628
Pattern Learning and Matching (PaLM)	0.4429
Extreme Gradient Boosting (XGBoost)	0.4363
Random Forest	0.4242
Ensemble Classifier (Voting)	0.4194
Decision Tree	0.4175
Neural Network	0.3558
Word2Vec Embedding	0.1763

A.5.2 BERT Model Performance

Based on the model selection results, I chose the fine-tuned pre-trained Chinese BERT with the Whole Word Masking model in the main analysis. The Chinese BERT model is a state-of-the-art deep learning model based on the Transformer architecture and pre-trained on a massive amount of Chinese text data. BERT learns contextualized representations of words by considering the entire sentence, capturing complex relationships between words. For text classification, the model takes the input text, tokenizes it, and passes it through its layers to produce embeddings. The embeddings are then used for topic classification tasks. The in-sample accuracy rate of the BERT model is 0.96 and the out-sample performance is presented in Table A.14.

Table A.14: Out-sample Five-fold Cross-Validation

	ADS	BET	COL	CRI	ESX	FOR	GOV	LCT	OTH	Macro
Precision	0.73	0.74	0.70	0.52	0.77	0.89	0.68	0.65	0.70	0.71
Recall	0.81	0.72	0.75	0.50	0.76	0.87	0.70	0.64	0.70	0.72
F1	0.77	0.73	0.72	0.51	0.76	0.88	0.69	0.65	0.70	0.71

Note: ADS: Advertisement. BET: Business. COL: Collective Action. CRI: Government Criticism. ESX: Entertainment. FOR: Foreign Events. GOV: Government (Others). LCT: Cultures. OTH: Others.

However, one concern arises regarding potential imbalances within the nine categories. To address this issue, I calculate category-specific weights to adjust for variations in category sizes. This approach allows for a more accurate assessment of the balanced performance of the BERT model. The results indicate a balanced precision of 0.71, a balanced recall of 0.73, and a balanced macro F1 score of 0.72.

Furthermore, I combine the three highly political categories, as well as the remaining six non-political categories, transforming the classification task into a binary one. When determining whether an article falls into the highly political category or not, the BERT model excels with a balanced precision of 0.83, a balanced recall of 0.83, and a balanced macro F1 score of 0.83.

A.5.3 Logistic Regression Model with Ridge Estimator

As a robustness check, I use the second-best-performing model, the multinomial logistic regression model with a ridge estimator, to re-run the classification task using one of the three data sources. I chose penalized regression models because the number of predictors (text) is much larger than the number of observations. Since I do not wish to drop predictors in the regularization process, the L2 ("ridge") penalty is preferable to the L1 ("LASSO") penalty. The training model is specified as:

$$y_{ij} = \alpha_j + \mathbf{DFM}_i \mathbf{fi_j} + \epsilon_{ij}$$

where y_{ij} is a binary variable that takes 1 if observation i belongs to topic category j and 0 otherwise. **DFM** is the document-feature matrix of the labeled data. **X** is a matrix of

additional predictors. $\mathbf{fi_j}$ is the matrix of ridge estimators for category j. Once the best matrices of ridge estimators, $\hat{\mathbf{fi}}_\mathbf{j}$, were found, I matched the unlabeled text corpus with the DFM of the labeled data. I then used the matched matrix and the best matrix of ridge estimators, $\hat{\mathbf{fi}}_\mathbf{j}$, to predict the unlabeled data.

Before the text analysis, all punctuation and stop words are removed and the Chinese text is segmented into individual tokens. Then, the segmented text was converted into a document-feature matrix. Words that appear less than 4 times were removed from the document-feature matrix.

Table A.15 shows that predictions are generally consistent with the main findings, with highly politically threatening content accounting for less than 40% of all censored articles. This confirms the theoretical expectation that moderate and non-political content accounts for the majority of all censored content.

Table A.15: Predicted Proportion of Censored Articles by Topic Category – Alternative Models

General Category	Specific Category	Logistical Regression (Ridge)
Highly Political	Collective Action	0.71%
	Govt Criticism	26.71%
	Other Govt-related	10.91%
	Total	38.33%
Moderately Political	Business	14.48%
	Foreign	3.89%
	Total	18.37%
Non-Political	Entertainment	19.38%
	Advertisement	7.79%
	Culture	12.74%
	Others	3.38%
	Total	43.29%

Notes: Data Source: WeChatScope, 15,872 censored articles.

Appendix B

Participatory Censorship in Authoritarian Regimes Appendix

B.1 Compliance with Ethical Principles of Human Subject Research

The two surveys conducted in this study followed all established principles of human subject research and was approved by the Institutional Review Board (IRB) at the researcher's home institution. Although the IRB exempted this study from a formal consent form, I still included a consent page and information sheet at the beginning of the survey. All participants were informed about the purpose, the length, and the format of the study. All participants need to click "I consent" on the information sheet page before they could proceed. They were allowed to opt-out of the study at any point of the survey. Incomplete survey responses were not recorded.

In the survey experiment (study 2), I showed respondents simulated social media posts that I adapted from Sina Weibo. Respondents are fully informed about how these posts are created from real posts, and therefore no deception is used.

All respondents were paid by the survey firm at its usual rate for their participation. The survey firm was paid by the researcher of this study. All participants were adults and none of them would be put in a disadvantageous position had they chosen not to participate.

Because this survey was conducted in China, an authoritarian regime, I paid extra caution to protect respondents' information and responses, so that they will not be negatively affected by the authority due to their participation in this study. I did not ask for personal information that could directly identify participants' identity, such as names, phone numbers, and email addresses. I stored all the responses at Qualtrics via an American institutional account. The study passed the information security review at the researcher's home institution.

B.2 Study 1: Sample and Weighting

Table B.1: Descriptive Statistics of the Original and Weighted Survey Sample (N=1,124)

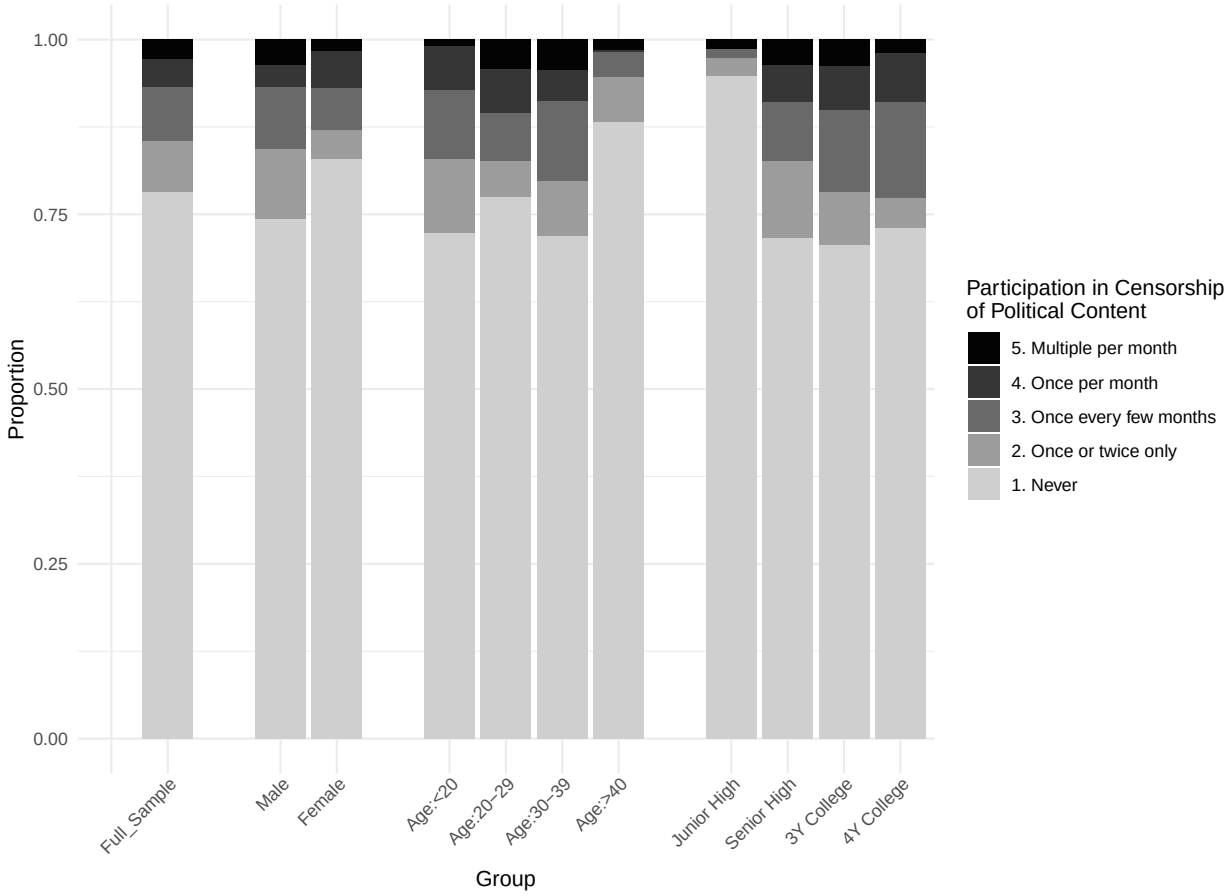
Sociodemographic Variables		Original Survey Sample	Weighted Survey Sample	China Internet Census
Gender	Female	44.9%	45.2%	47.3%
	Male	54.8%	54.7%	52.7%
Location	Rural	30.7%	29.7%	28.2%
	Urban	67.9%	69.7%	71.8%
Region	East	25.2%	29.3%	31.1%
	South & Central	35.2%	29.9%	28.2%
	North & Northeast	26.0%	21.0%	22.2%
	West	12.7%	18.9%	18.5%
Age	≤ 19	8.6%	22.0%	21.6%
	20-29	40.0%	26.7%	26.8%
	30-39	32.5%	22.2%	23.5%
	≥ 40	17.4%	28.6%	28.1%
Education	$\leq$ High School	26.1 %	77.4%	79.8%
	$\geq$ College	73.6%	21.4%	20.2%

Note: Data about Chinese Internet users are from *The 45th Statistical Report of Internet Development in China*, issued by China Internet Network Information Center in April 2020.

B.3 Study 1: Prevalence of Participation

B.3.1 Participation in Censorship of Specific Content Categories

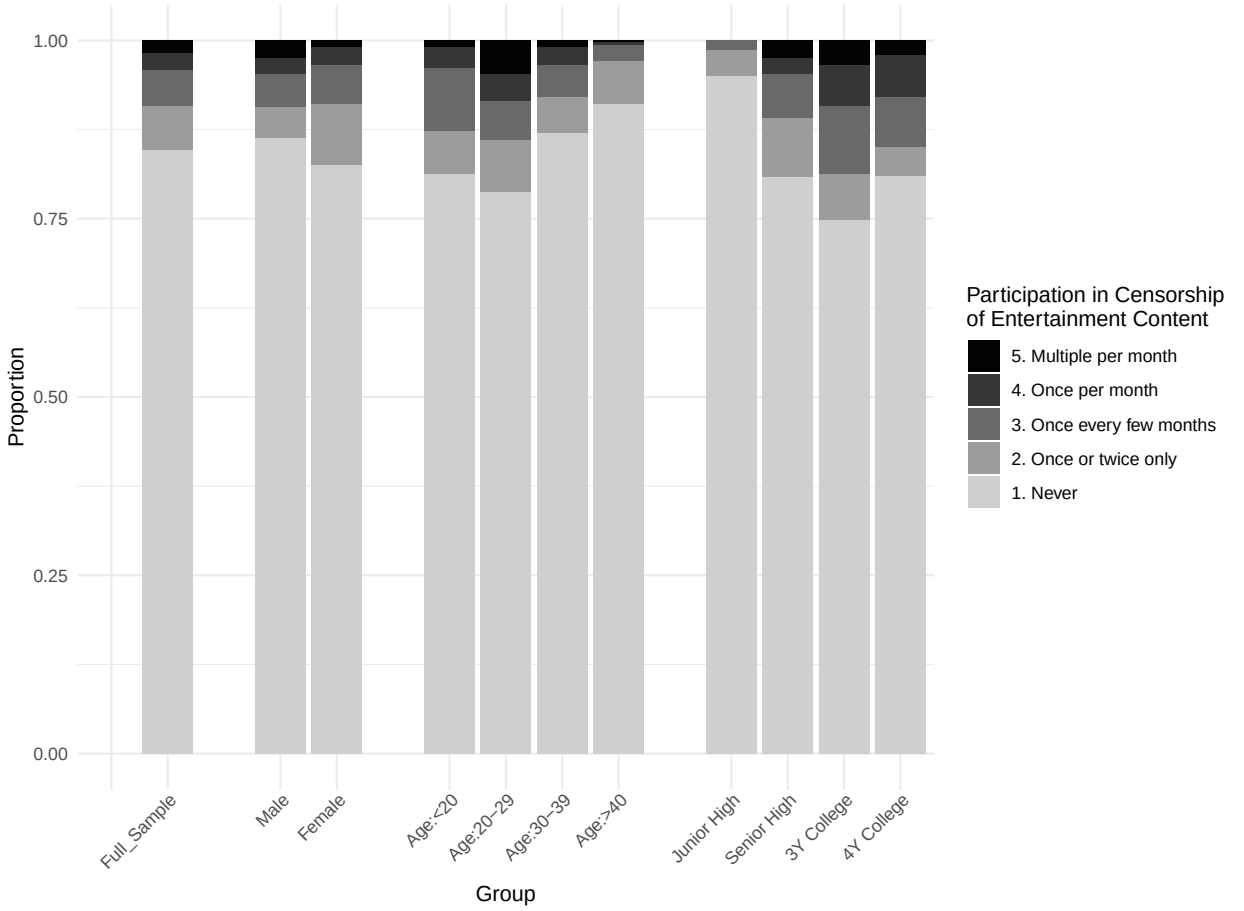
Figure B.1 reports the distribution of self-reported participation in the censorship of political content. In general, around 50% of the "participating respondents," or 25% of all respondents, self-report having participated in the censorship of political content. Men, younger generations, and the better-educated are significantly more likely to participate in the censorship of political content.



Note: All observations are weighted by gender, rural/urban location, region, age group, and education.

Figure B.1: Distribution of Self-Report Participation in Censorship of Political Content

Figure B.3 reports the distribution of self-reported participation in the censorship of entertainment content. Around one-third of the "participating respondents," or one-sixth of all respondents, self-reporting having participated in the censorship of entertainment content. In contrast to political content, females are more likely to report entertainment content than males. Consistent with the political content, younger and better-educated are more likely to report entertainment content.



Note: All observations are weighted by gender, rural/urban location, region, age group, and education.

Figure B.2: Distribution of Self-Report Participation in Censorship of Entertainment Content

B.3.2 Unweighted Sample

The unweighted sample shows a slightly higher proportion of respondents self-reporting participation in censorship.

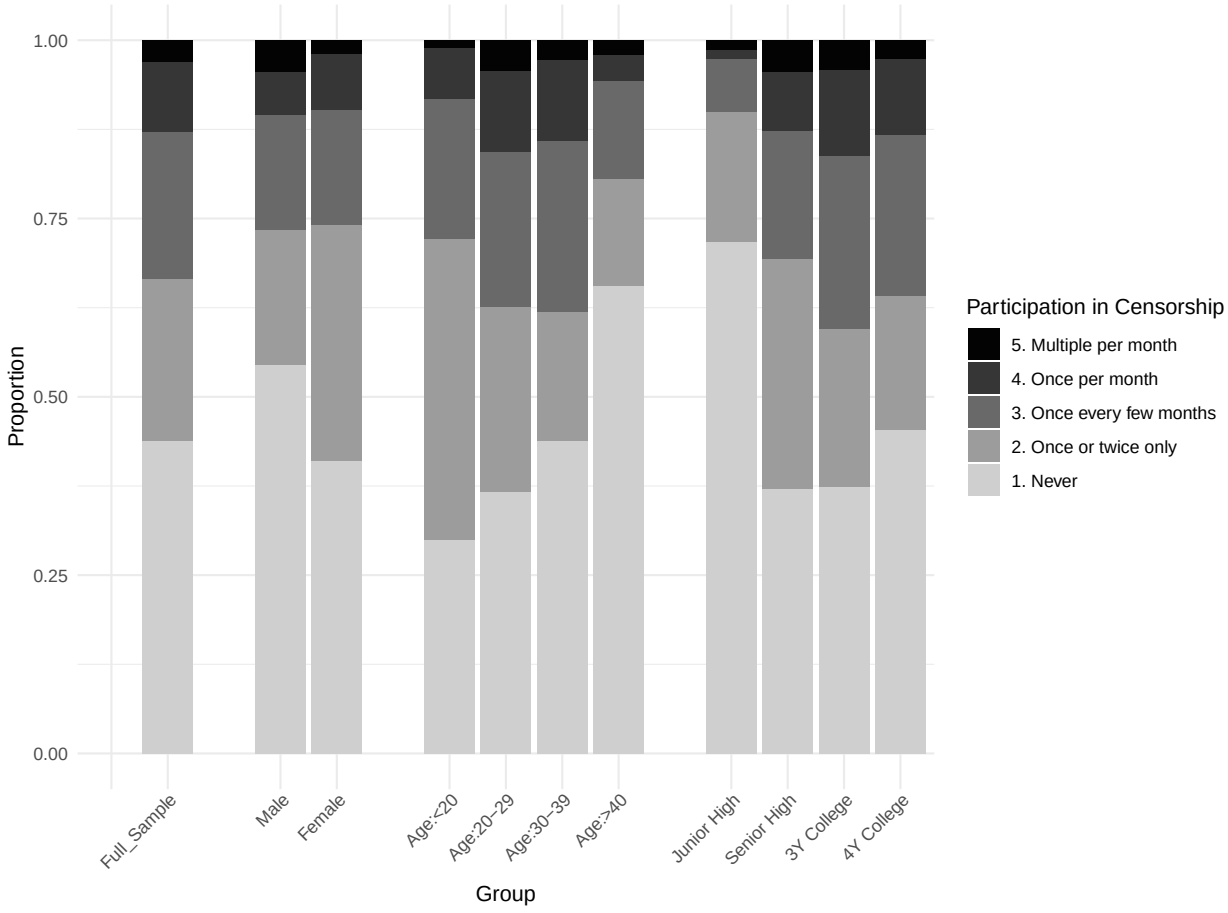


Figure B.3: Distribution of Self-Report Participation in Censorship: Unweighted Sample

B.4 Study 1: Correlation with Support

B.4.1 Main Analyses

Table B.2: Correlation between Participation in Censorship and Support for Censorship Using the Five Point Measure of Participation

	Support for Censorship		Support for Censorship of Political Content		Support for Censorship of Non-Political Content	
	(1)	(2)	(3)	(4)	(5)	(6)
Participation	0.099***	0.084***	0.085***	0.105***	0.007	0.027
	(0.028)	(0.028)	(0.030)	(0.030)	(0.033)	(0.033)
Female	0.179***	0.198***	−0.031	−0.046	0.012	0.0001
	(0.064)	(0.063)	(0.066)	(0.067)	(0.073)	(0.073)
Age Group	0.156***	0.127***	0.052*	0.031	0.104***	0.084***
	(0.027)	(0.027)	(0.028)	(0.028)	(0.031)	(0.031)
Education	−0.062**	−0.068**	0.060*	0.059*	0.036	0.035
	(0.030)	(0.031)	(0.031)	(0.032)	(0.034)	(0.035)
Urban	0.214***	0.225***	0.375***	0.338***	0.304***	0.286***
	(0.067)	(0.067)	(0.070)	(0.071)	(0.077)	(0.078)
Party Member		0.391***		−0.001		0.120
		(0.097)		(0.103)		(0.113)
Pol. Ideology		−0.003		−0.112***		−0.138***
		(0.025)		(0.026)		(0.029)
Econ. Ideology		0.202***		0.036		0.031
		(0.030)		(0.032)		(0.035)
Pol. Interest		−0.026		0.070***		0.093***
		(0.022)		(0.023)		(0.025)
Constant	2.616***	2.067***	2.549***	2.583***	2.367***	2.387***
	(0.131)	(0.182)	(0.136)	(0.191)	(0.150)	(0.211)
Weighted Sample	✓	✓	✓	✓	✓	✓
N	1,088	1,071	1,084	1,066	1,086	1,068
Adjusted R^2	0.048	0.106	0.046	0.074	0.034	0.070

Notes: Dependent variables are indicated in column headings and are measured on a five-point Likert scale. Standard errors in parentheses. Participation in censorship is measured on a five-point scale: never participated, once or twice only, once per few months, once per month, and multiple times per month.
*p < .1; **p < .05; ***p < .01

Table B.3: Correlation between Specific Types of Participation in Censorship and Support for Censorship Using the Five Point Measure of Participation

	Support for Censorship		Support for Censorship of Political Content		Support for Censorship of Non-Political Content	
	(1)	(2)	(3)	(4)	(5)	(6)
Participation (Political)	0.059**		0.121***		0.004	
	(0.030)		(0.032)		(0.035)	
Participation (NonPolitical)		0.053		0.079**		−0.001
		(0.036)		(0.037)		(0.043)
Female	0.197***	0.187***	−0.036	−0.056	−0.003	−0.004
	(0.063)	(0.063)	(0.067)	(0.067)	(0.073)	(0.073)
Age Group	0.110***	0.108***	0.017	0.010	0.077**	0.076**
	(0.026)	(0.026)	(0.027)	(0.027)	(0.030)	(0.030)
Education	−0.063**	−0.061**	0.059*	0.067**	0.038	0.039
	(0.031)	(0.031)	(0.032)	(0.032)	(0.035)	(0.035)
Urban	0.223***	0.222***	0.332***	0.330***	0.286***	0.286***
	(0.068)	(0.068)	(0.071)	(0.071)	(0.078)	(0.078)
Party Member	0.402***	0.411***	−0.001	0.009	0.129	0.131
	(0.097)	(0.097)	(0.103)	(0.104)	(0.113)	(0.113)
Pol. Ideology	0.001	0.004	−0.107***	−0.104***	−0.137***	−0.137***
	(0.025)	(0.025)	(0.026)	(0.026)	(0.029)	(0.029)
Econ. Ideology	0.207***	0.207***	0.041	0.040	0.033	0.033
	(0.030)	(0.030)	(0.032)	(0.032)	(0.035)	(0.035)
Pol. Interest	−0.026	−0.024	0.070***	0.074***	0.093***	0.093***
	(0.022)	(0.022)	(0.023)	(0.023)	(0.025)	(0.025)
Constant	2.151***	2.159***	2.618***	2.682***	2.435***	2.443***
	(0.178)	(0.181)	(0.187)	(0.190)	(0.207)	(0.211)
Weighted Sample	✓	✓	✓	✓	✓	✓
N	1,071	1,071	1,066	1,066	1,068	1,068
Adjusted R^2	0.102	0.100	0.076	0.067	0.069	0.069

Notes: Dependent variables are indicated in column headings and are measured on a five-point Likert scale. Standard errors in parentheses. The independent variables are participation in political censorship and participation in censorship of entertainment and cultural content. Both independent variables are measured on a five-point scale: never participated, once or twice only, once per few months, once per month, and multiple times per month.

*p < .1; **p < .05; ***p < .01

B.4.2 Robustness Checks

First, I transform the independent variable into a binary measure of participation. If the respondent has never participated before, I code it as 0, otherwise, I code it as 1. Table B.14 shows the results using the binary measure of participation to re-run the analyses. As demonstrated in the table, the relationship between participation and support for censorship remains the same using the binary measure.

Table B.4: Correlation between Participation in Censorship and Support for Censorship Using Binary Measure of Participation

	Support for Censorship		Support for Censorship of Political Content		Support for Censorship of Non-Political Content	
	(1)	(2)	(3)	(4)	(5)	(6)
Participate (Binary)	0.184***	0.189***	0.184***	0.214***	−0.135*	−0.103
	(0.064)	(0.063)	(0.066)	(0.067)	(0.073)	(0.073)
Female	0.165***	0.191***	−0.040	−0.053	0.008	−0.008
	(0.063)	(0.063)	(0.066)	(0.067)	(0.073)	(0.073)
Age Group	0.155***	0.133***	0.056**	0.035	0.081**	0.059*
	(0.028)	(0.028)	(0.028)	(0.029)	(0.032)	(0.032)
Education	−0.056*	−0.065**	0.064**	0.064**	0.045	0.044
	(0.030)	(0.030)	(0.031)	(0.032)	(0.034)	(0.035)
Urban	0.225***	0.240***	0.386***	0.355***	0.297***	0.280***
	(0.067)	(0.068)	(0.070)	(0.071)	(0.077)	(0.078)
Party Member		0.421***		0.030		0.133
		(0.097)		(0.103)		(0.112)
Pol. Ideology		−0.007		−0.117***		−0.132***
		(0.025)		(0.027)		(0.029)
Econ. Ideology		0.204***		0.038		0.035
		(0.030)		(0.032)		(0.035)
Pol. Interest		−0.031		0.064***		0.095***
		(0.022)		(0.023)		(0.025)
Constant	2.697***	2.123***	2.594***	2.669***	2.498***	2.504***
	(0.125)	(0.177)	(0.129)	(0.186)	(0.143)	(0.205)
Weighted Sample	✓	✓	✓	✓	✓	✓
N	1,088	1,071	1,084	1,066	1,086	1,068
Adjusted R^2	0.019	0.106	0.046	0.088	0.020	0.088

Notes: Dependent variables are indicated in column headings and are measured on a five-point Likert scale. Standard errors in parentheses. Participation is a binary variable indicating whether the respondent has participated before.
*p < .1; **p < .05; ***p < .01

B.4.3 Additional Mechanisms

An additional mechanism that I tested is the increase in the perceived benefit of censorship. Specifically, participation allows ordinary users to report content they disapprove of, thereby increasing their perceived benefit. When censorship is solely a top-down process imposed upon ordinary users, they are more likely to have cynical views of the censorship apparatus and perceive themselves as victims of censorship. Conversely, because of the increased perceived empowerment, individuals are more likely to view government censorship activities as enforcing their censorship preferences. From the perspective of ordinary users, their participation redefines the government's role as an arbitrator of public demand on the Internet, rather than a manipulator of public opinion. Hence, they are more likely to view censorship as a tool they can use to suppress political opponents, increasing their support for censorship.

Hypothesis: As individuals participate more in the censorship process, they are more likely to believe that censorship benefits ordinary citizens such as themselves, which subsequently leads to greater levels of support for government censorship.

To examine whether participation increases the perceived benefit of censorship, the survey asked respondents about whether ordinary people are the victims or the beneficiaries of the current censorship apparatus. Using OLS regression models with all relevant covariates and adjusted by sample weights, I find no evidence to support the perceived benefit mechanism ($\beta = -0.077$, $p = 0.191$). The insignificant results may be due to the lack of variation in responses, which are concentrated on the middle choices. This suggests that respondents may not have a strong opinion on the perceived benefit question.

B.5 Study 2: Experimental Design & Randomization Check

B.5.1 Simulated Social Media Posts

Figure B.4 shows an example of the simulated social media posts. In the control groups (Upper Panel), there are three buttons under each post: "Like," "Share," and "Comment." In both treatment groups (Lower Panel), there are four buttons, a "Report" button (the button on the right) in addition to the three in the control group.

Figure B.4: Simulated Social Media Posts

B.5.2 Question Wording

Table B.5: Measurement of Main Outcome Variables

Hypothesis	Survey Items	Expectation
Censorship Support	Do you agree or disagree: The government should actively control the Internet and remove content that it deems inappropriate.	+
	Do you agree or disagree: The government should actively control online discussions on **government policies and party leadership**, and remove content that it deems inappropriate.	+
	Do you agree or disagree: The government should actively control online discussions on **entertainment stars and popular culture**, and remove content that it deems inappropriate.	+
Regime Support	How satisfied are you with the overall situation in China right now?	+
	Both the central and the local government of all levels always works for the people and serves their needs.	+
	I completely trust both the central and the local government.	+

B.5.3 Balance Table & Randomization Check

Table B.6: Balance Table (Group Mean & *F*-test)

	Control	Treatment 1	Treatment 2	*p*-value
Female	2.76	2.77	2.75	0.465
Age Group	2.99	3.04	3.00	0.830
Education	0.48	0.48	0.50	0.686
Party Member	0.83	0.80	0.82	0.216
Economic Ideology	3.79	3.80	3.81	0.662
Nationalism	0.16	0.17	0.14	0.955
Political Interest	3.54	3.59	3.56	0.513
Social Media Usage	2.85	2.87	2.85	0.759
Foreign Connection	3.72	3.71	3.73	0.781

Table B.7: Randomization Check: Using Covariates to Predict Treatment

	Group Assignment	
	Treatment 1 − Control	Treatment 2 − Treatment 1
Female	−0.005	0.017
	(0.020)	(0.020)
Age Group	−0.002	0.00003
	(0.010)	(0.010)
Education	0.016	−0.014
	(0.012)	(0.012)
Party Member	0.008	−0.048
	(0.028)	(0.029)
Economic Ideology	−0.001	0.002
	(0.011)	(0.011)
Region	0.002	−0.002
	(0.006)	(0.006)
Nationalism	0.003	−0.003
	(0.008)	(0.008)
Political Interest	0.012	−0.005
	(0.009)	(0.009)
Social Media Usage	−0.006	0.008
	(0.010)	(0.010)
Foreign Connection	−0.022	0.013
	(0.012)	(0.012)

$^{*}p < .05$; $^{**}p < .01$

B.6 Study 2: Analyses

B.6.1 Overall Results

Comparing the Control Group and Treatment Group 1

Table B.8: The Effect of Providing the Opportunity to Participate on Support for Censorship

	Support for Censorship		Support for Censorship of Political Content		Support for Censorship of Non-Political Content	
	(1)	(2)	(3)	(4)	(5)	(6)
Control Group	0.067*	0.065*	0.066*	0.071*	0.028	0.025
	(0.036)	(0.035)	(0.037)	(0.036)	(0.038)	(0.038)
Female		−0.076**		0.013		−0.015
		(0.036)		(0.037)		(0.039)
Age Group		0.115***		0.088***		0.078***
		(0.018)		(0.019)		(0.020)
Education		0.005		0.020		0.001
		(0.021)		(0.022)		(0.023)
Party Member		0.075		0.033		0.055
		(0.050)		(0.051)		(0.053)
Ideology		0.299***		0.285***		0.275***
		(0.020)		(0.021)		(0.022)
Nationalism		−0.004		0.018		−0.024*
		(0.013)		(0.014)		(0.014)
Political Interest		0.039**		0.013		0.045**
		(0.016)		(0.017)		(0.018)
Social Media Usage		−0.002		−0.019		−0.030
		(0.018)		(0.018)		(0.019)
Foreign Connection		−0.059***		−0.040*		−0.030
		(0.021)		(0.022)		(0.023)
Constant	3.491***	1.980***	3.468***	2.069***	3.679***	2.452***
	(0.026)	(0.137)	(0.026)	(0.142)	(0.027)	(0.147)
N	2,664	2,504	2,668	2,507	2,662	2,501
Adjusted R^2	0.001	0.119	0.001	0.092	−0.0002	0.081

Notes: Dependent variables are indicated in column headings and are measured on a five-point Likert scale. Standard errors in parentheses. Only the blank control and control groups are included in the analyses.
*p < .1; **p < .05; ***p < .01

Comparing Treatment Groups 1 and 2

Table B.9: Intention-To-Treat Effect of the Encouragement Treatment

	Support for Censorship		Support for Censorship of Political Content		Support for Censorship of Non-Political Content	
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment Group	0.086**	0.081**	0.068**	0.061*	0.062*	0.064*
	(0.034)	(0.033)	(0.034)	(0.034)	(0.037)	(0.037)
Female		−0.027		0.046		−0.025
		(0.034)		(0.035)		(0.037)
Age Group		0.098***		0.068***		0.087***
		(0.017)		(0.018)		(0.019)
Education		−0.014		0.023		−0.004
		(0.020)		(0.020)		(0.022)
Party Member		0.063		0.042		0.011
		(0.048)		(0.049)		(0.053)
Ideology		0.240***		0.228***		0.236***
		(0.019)		(0.019)		(0.021)
Nationalism		−0.018		0.005		−0.024*
		(0.013)		(0.013)		(0.014)
Political Interest		0.061***		0.047***		0.072***
		(0.015)		(0.016)		(0.017)
Social Media Usage		−0.003		−0.032*		−0.038**
		(0.017)		(0.017)		(0.019)
Foreign Connection		−0.038*		−0.040*		−0.045**
		(0.020)		(0.021)		(0.022)
Constant	3.559***	2.278***	3.534***	2.327***	3.707***	2.455***
	(0.024)	(0.128)	(0.024)	(0.130)	(0.026)	(0.141)
N	2,647	2,493	2,653	2,499	2,645	2,492
Adjusted R^2	0.002	0.097	0.001	0.075	0.001	0.082

Notes: Dependent variables are indicated in column headings and are measured on a five-point Likert scale. Standard errors in parentheses. Only the treatment and control groups are included in the analyses.
*p < .1; **p < .05; ***p < .01

B.6.2 Instrumental Variable Analysis

Main Analysis

Table B.10: Complier Average Causal Effects (CACE) of Participating in Censorship on Support for Censorship: Main Analysis

	Support for Censorship		Support for Censorship of Political Content		Support for Censorship of Non-Political Content	
	(1)	(2)	(3)	(4)	(5)	(6)
Report Click	0.228***	0.219***	0.202***	0.199***	0.130**	0.124**
	(0.054)	(0.052)	(0.055)	(0.054)	(0.058)	(0.056)
Female		−0.046		0.030		−0.015
		(0.028)		(0.029)		(0.031)
Age		0.136***		0.104***		0.103***
		(0.015)		(0.015)		(0.016)
Education		0.004		0.023		0.002
		(0.017)		(0.017)		(0.018)
Party Member		0.082**		0.055		0.041
		(0.040)		(0.042)		(0.043)
Ideology		0.277***		0.263***		0.261***
		(0.016)		(0.017)		(0.017)
Nationalism		0.001		0.017		−0.021*
		(0.011)		(0.011)		(0.012)
Political Interest		0.035***		0.019		0.053***
		(0.013)		(0.013)		(0.014)
Social Media		0.005		−0.021		−0.039**
		(0.014)		(0.015)		(0.015)
Foreign		−0.048***		−0.046***		−0.042**
		(0.017)		(0.017)		(0.018)
Constant	3.484***	1.954***	3.463***	2.082***	3.673***	2.386***
	(0.024)	(0.110)	(0.025)	(0.114)	(0.025)	(0.119)
N	3,990	3,764	3,997	3,770	3,989	3,763

Notes: Report click is a binary variable indicating whether the respondents have clicked any of the "Report" buttons on the simulated social media page.
*p < .1; **p < .05; ***p < .01

Robustness Check

To check the robustness of the treatment effect, I use an alternative measurement of participation in censorship: the number of times the respondents clicked a "Report" button. As shown in Table B.11, consistent with the main analyses, additional clicking of the "Report" buttons induced by the treatments significantly increases support for censorship.

Table B.11: Complier Average Causal Effects (CACE) of Participating in Censorship on Support for Censorship: Alternative Measurement

	Support for Censorship		Support for Censorship of Political Content		Support for Censorship of Non-Political Content	
	(1)	(2)	(3)	(4)	(5)	(6)
Report Click #	0.080***	0.077***	0.071***	0.070***	0.046**	0.044**
	(0.019)	(0.018)	(0.020)	(0.019)	(0.020)	(0.020)
Female		−0.042		0.033		−0.013
		(0.029)		(0.030)		(0.031)
Age		0.141***		0.109***		0.106***
		(0.015)		(0.016)		(0.017)
Education		0.007		0.026		0.004
		(0.017)		(0.017)		(0.018)
Party Member		0.082**		0.055		0.041
		(0.041)		(0.042)		(0.044)
Ideology		0.278***		0.263***		0.262***
		(0.016)		(0.017)		(0.017)
Nationalism		0.004		0.020*		−0.020*
		(0.011)		(0.011)		(0.012)
Political Interest		0.036***		0.019		0.053***
		(0.013)		(0.013)		(0.014)
Social Media		0.001		−0.024		−0.041***
		(0.014)		(0.015)		(0.015)
Foreign		−0.050***		−0.048***		−0.044**
		(0.017)		(0.018)		(0.018)
Constant	3.484***	1.935***	3.463***	2.066***	3.673***	2.376***
	(0.024)	(0.112)	(0.025)	(0.116)	(0.026)	(0.121)
N	3,990	3,764	3,997	3,770	3,989	3,763

Notes: Report click number is the number of the "Report" buttons that the respondents clicked on the simulated social media page. All individual survey items were measured on a five-point scale.
LATE = Local Average Treatment Effect
*p < .1; **p < .05; ***p < .01

I further subset the data to only include the two treatment groups and re-run the instrumental variable analyses. The results will represent the effect of participation induced by the encouragement treatment on support for censorship.

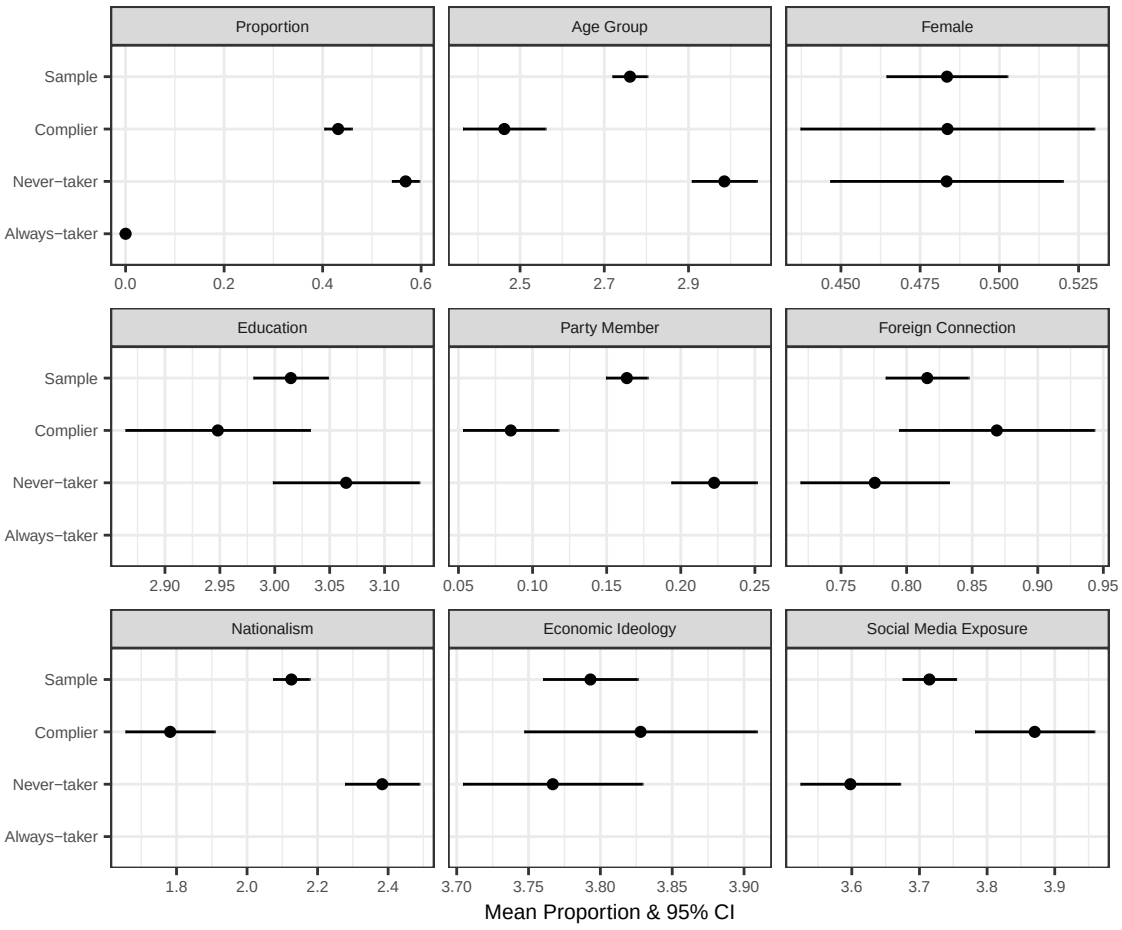
Table B.12: Complier Average Causal Effects (CACE) of Participating in Censorship on Support for Censorship: Two-Group Subset

	Support for Censorship		Support for Censorship of Political Content		Support for Censorship of Non-Political Content	
	(1)	(2)	(3)	(4)	(5)	(6)
Report Click	0.417**	0.402**	0.330*	0.305*	0.299*	0.319*
	(0.169)	(0.168)	(0.169)	(0.170)	(0.180)	(0.183)
Female		−0.030		0.044		−0.026
		(0.035)		(0.035)		(0.038)
Age		0.145***		0.103***		0.124***
		(0.026)		(0.027)		(0.029)
Education		−0.002		0.032		0.006
		(0.021)		(0.021)		(0.023)
Party Member		0.108**		0.076		0.046
		(0.053)		(0.053)		(0.058)
Ideology		0.235***		0.224***		0.232***
		(0.019)		(0.020)		(0.021)
Nationalism		0.005		0.022		−0.005
		(0.016)		(0.016)		(0.018)
Political Interest		0.053***		0.041**		0.066***
		(0.016)		(0.016)		(0.017)
Social Media		−0.010		−0.037**		−0.044**
		(0.018)		(0.018)		(0.019)
Foreign		−0.054**		−0.052**		−0.057**
		(0.022)		(0.022)		(0.024)
Constant	3.379***	1.977***	3.392***	2.097***	3.579***	2.216***
	(0.092)	(0.193)	(0.092)	(0.195)	(0.098)	(0.211)
N	2,647	2,493	2,653	2,499	2,645	2,492

Notes: Report click is a binary variable indicating whether the respondents have clicked any of the "Report" buttons on the simulated social media page.
*p < .1; **p < .05; ***p < .01

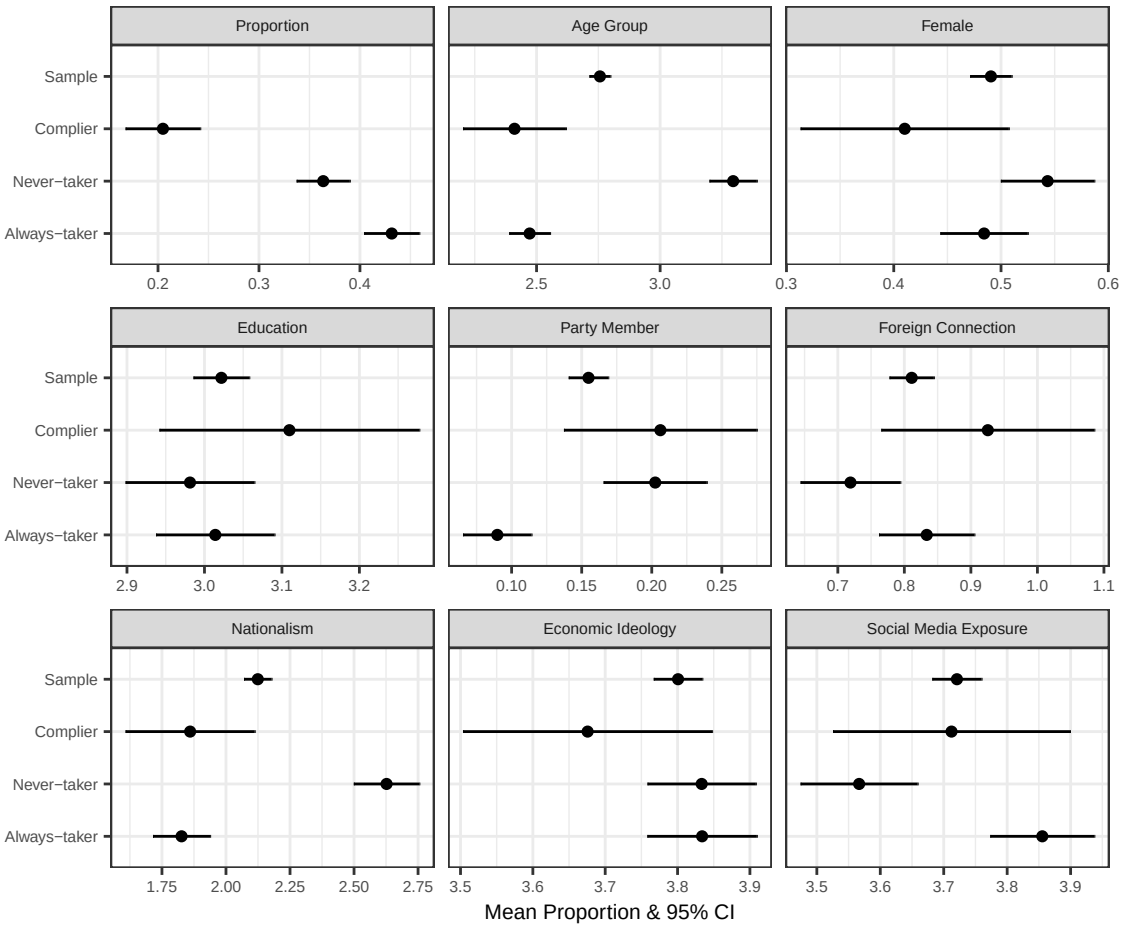
B.6.3 Profiling Compliers

Following Marbach and Hangartner (2020), I plot out the characteristics of compliers, always-takers, and never-takers, assuming there are no defiers. Figure B.5 uses the control group and treatment group 1. Because respondents in the control group cannot participate, there are no always-takers. Figure B.6 uses the two treatment groups. Overall, individuals who clicked the “Report" buttons tended to be younger and more familiar with social media. In contrast, the non-compliers, those who never click the “Report” buttons, tended to be older, nationalists with few foreign connections and limited social media exposure.



Note: The first panel indicates the estimated proportion of compliers, never-takers, and always-takers. The remaining eight panels demonstrate the estimated group means of the full sample, compliers, never-takers, and always-takers across eight different pre-treatment covariates. CI = Confidence Interval

Figure B.5: Profiling Compliers and Non-Compliers Using Control and Treatment 1



Note: The first panel indicates the estimated proportion of compliers and never-takers. The remaining eight panels demonstrate the estimated group means of the full sample, compliers, and never-takers across eight different pre-treatment covariates. CI = Confidence Interval

Figure B.6: Profiling Compliers, Never-Takers, and Always-Takers Using Treatment Groups 1 & 2

B.6.4 Regime Support

Consistent with the findings in Study 1, I do not find significant effects of participation treatment on regime supports. This provides additional support for the cognitive dissonance theory and the system justification theory as explained in the main text in section 3.4.

Comparing the Control Group and Treatment Group 1

Table B.13: The Effect of Providing the Opportunity to Participate on Regime Support

	Regime Satisfaction		Regime Assessment		Regime Trust	
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment Group 1	0.011	0.026	0.002	0.010	0.031	0.053*
	(0.032)	(0.032)	(0.032)	(0.031)	(0.033)	(0.032)
Constant	4.066***	3.180***	4.046***	3.344***	4.023***	3.060***
	(0.023)	(0.123)	(0.022)	(0.120)	(0.023)	(0.125)
Covariates		✓		✓		✓
N	2,636	2,477	2,636	2,476	2,645	2,487
Adjusted R^2	−0.0003	0.089	−0.0004	0.104	−0.00005	0.091

Notes: *p < .1; **p < .05; ***p < .01

Comparing Treatment Groups 1 & 2

Table B.14: Intention-To-Treat Effect of Encouragement Treatment on Regime Support

	Regime Satisfaction		Regime Assessment		Regime Trust	
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment Group 2	0.056*	0.041	0.051	0.039	0.029	0.012
	(0.032)	(0.032)	(0.031)	(0.031)	(0.032)	(0.031)
Constant	4.076***	3.504***	4.048***	3.451***	4.053***	3.324***
	(0.023)	(0.122)	(0.022)	(0.119)	(0.023)	(0.121)
Covariates		✓		✓		✓
N	2,616	2,465	2,618	2,466	2,628	2,479
Adjusted R^2	0.001	0.074	0.001	0.096	−0.0001	0.069

Notes: $^{*}p < .1$; $^{**}p < .05$; $^{***}p < .01$

Appendix C

How Chinese Censorship Allows Public Discourse on Democracies but Not Their Institutions Appendix

C.1 Topic Keywords

C.1.1 Democratic Institutions

1. Bureaucracy & Federalism
   Federalism, Federal, Secretary, Minister, Department, Federal Bureau of Investigation, Department of Justice, Department of Defense, Central Intelligence Agency

2. Diaspora Political Participation
   Asian, Chinese, Participate, Alliance, Voice, Civic, Citizens, Communities, Candidate, Race, Andrew Yang, Candidate

3. Elections
   Candidate, Running for, Election, General Election, Elected, Presidential Election, Voting, Nomination, Polling, Poll, Ballet, Campaign, Vote Counting, Win, Reelection, Electoral College, Swing

4. Executive (Asia)
   Prime Minister, President, Cabinet, India, Modi, BJP, Singh, Japanese Government, LDP, Shinzo Abe, Fumio Kishida, Yoshihide Suga, Blue House, Moon Jae-in, Yoon Suk Yeol, Park Geun-hye

5. Executive (Europe)
   Prime Minister, President, Cabinet, Minister, United Kingdom, France, Germany, European Union, Italy, Rishi Sunak, Liz Truss, Borris Johnson, Conservative Party, Emmanuel Macron, Marine Le Pen, Angela Merkel, Olaf Scholz

6. Executive (US)
   President, US Government, Federal Government, White House, Donald Trump, Joe Biden, Barack Obama, Bill Clinton, George Bush

7. Inter-branch Checks & Balances
   Impeachment, Constitution, Nomination, Override, Separation of Power, Checks and Balances, Judicial Independence, Power, Legislature, Executive, Judiciary, Veto

8. Legislature
   Congress, Parliament, Senate, House of Representatives, Legislator, Congressperson, Senator, Committee, Bill, Law, Act, Vote

9. Media
   Media, Free Speech, Censor, Ban, Journalist, Investigation, CNN, NYT, WSJ, CBS, Fox, Twitter, Social Media, Fake News, Speech

10. Judiciary
    Supreme Court, Justices, Judge, Court, Ruling, Law, Appeal, Prosecutor, Constitution, Case

11. Political Liberalism
    Democracy, Conservatism, Liberalism, Veto, Human Rights, Society, Western, Independent, Institutions, Natural Law, Common Law, Thomas Hobbs, Jean-Jacques Rousseau, Michel Foucault

12. Economic Liberalism
    Economics, Private Property, Property Rights, Market, Free Market, Property, Capitalism, Productive Forces, Adam Smith, John Maynard Keynes, Friedrich Hayek

C.1.2 Socioeconomic Conditions

1. Protests (BLM)
   Protest, Protesters, Racism, White, Black, BLM, George Floyd, Violence, Demonstrations, Racism, Riot, Police, Discrimination
2. COVID-19 Pandemic
   COVID, Pandemic, Quarantine, Positive, Pfizer, Cases, Symptom, Virus, Vaccine, Epidemic, Testing, Death, Spread, Mask, Mutation, Coronavirus, WHO, Asymptomatic
3. Crime & Gun Violence
   Police, Criminal, Suspect, Gun, Gun Man, Shooting, Victim, Arrest, Fire, Robbery, Death, Investigation, Drugs, Illegal, Carry
4. Cybersecurity
   Cybersecurity, Hacker, Internet, Loophole, Cyber Attack, Data, Leak, Steal, Information, Attacker
5. Diaspora Livelihood (US)
   Chinese, Chinese Community, Chinese American, Information, Dining, Fun, Play, Business, Service, Life, Seattle, New York, Los Angeles, Texas, San Francisco, California
6. Diaspora Livelihood (Others)
   Chinese, Chinese Community, Information, Dining, Fun, Play, Business, Service, Life, Toronto, Vancouver, Sydney, Melbourne, Australia, New Zealand, Madrid, Spain, Germany
7. Disaster
   Disaster, Victim, Dead, Death, Fire, Wildfire, Flood, Explosion, Attack, Earthquake, Typhoon, Accident
8. Economy
   Economy, GDP, Growth, Debt, Crisis, Currency, Recession, Population, IMF, Per Capita, Development, Trade, Import, Export, Income, Inflation
9. Education
   University, College, High School, Exam, Study Abroad, Graduate, Undergraduate, Score, Admission, Major, Ivy League, Harvard, Doctor, Professor, Student

10. Emigration to Democracies
    Immigration, Emigration, Immigrant, Visa, Citizenship and Immigration Services, Green Card, Naturalization, Refugee, Deport, Application, Citizenship

11. Energy
    Resource, Energy, Oil, Clean Energy, Liquefied, Rare Earth, Gasoline, Natural Gas, Coal, Solar Energy, Energy Crisis

12. Environment
    Pollution, Temperature, Climate, Climate Change, Forest, Emission, Weather, Waste, Protect, Global, Human, Nuclear, Carbon

13. Housing
    House, Housing Price, Real Estate, Housing, Buying House, Condo, Rent, Apartment

14. Healthcare
    Treatment, Gene, Pharmaceutical, Cancer, Healthcare, Diagnosis, Immunity, Clinical Trial, Insurance

15. Stock Market
    Rise, Fall, S&P 500, Interest Rate, Mutual Fund, Fund, Invest, Market, Index, Dollar, US Stock, Capital, Finance, Bank, Inflation, Oil Price

16. Technology (AI)
    AI, Artificial Intelligence, Computer, Machine Learning, Quantum, Facial Recognition, Algorithm, Robot, Science

17. Technology (Chips)
    Chips, Semiconductor, TSMC, Qualcomm, Intel, Foxconn

18. Technology (EV)
    Tesla, EV, Electric, Car, Charging, Fast Charging, Driving, Auto

19. Technology (Big Tech Firms)
    Google, Microsoft, Amazon, Apple, Facebook, Internet, Giant, iPhone, iPod, CEO

C.2 Results Using Full Censorship Data

In the main paper, I dropped the articles directly related to China because topics related to domestic politics might be subject to a different set of censorship logic. In this section, I demonstrate that the main findings in the paper remain substantially unchanged when including those articles related to China.

In Figure C.1, I replicate the main analysis comparing the censorship rate by topic categories and stance. Consistent with the main findings, content related to democratic institutions is around three times more likely to be censored than content related to socioeconomic conditions in democracies. Moreover, stance and sentiment appear to play minimal role in determining censorship incidence.

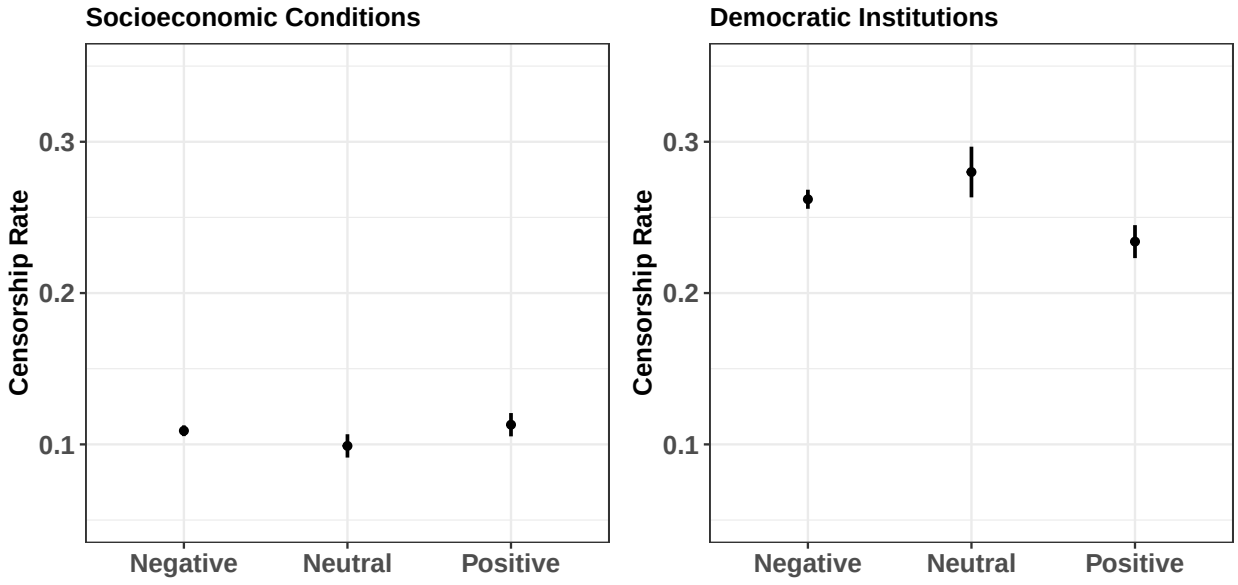


Figure C.1: Censorship Rate of Democratic Institutions and Socioeconomic Conditions by Stance: Full Censorship Data

Figure C.2 further disaggregates the censorship rate by specific topic categories. The pattern is generally consistent with both the main analysis, as well as the in Figure C.3. Overall, topics related to democratic institutions are more likely to be censored than topics related to socioeconomic conditions in democracies.

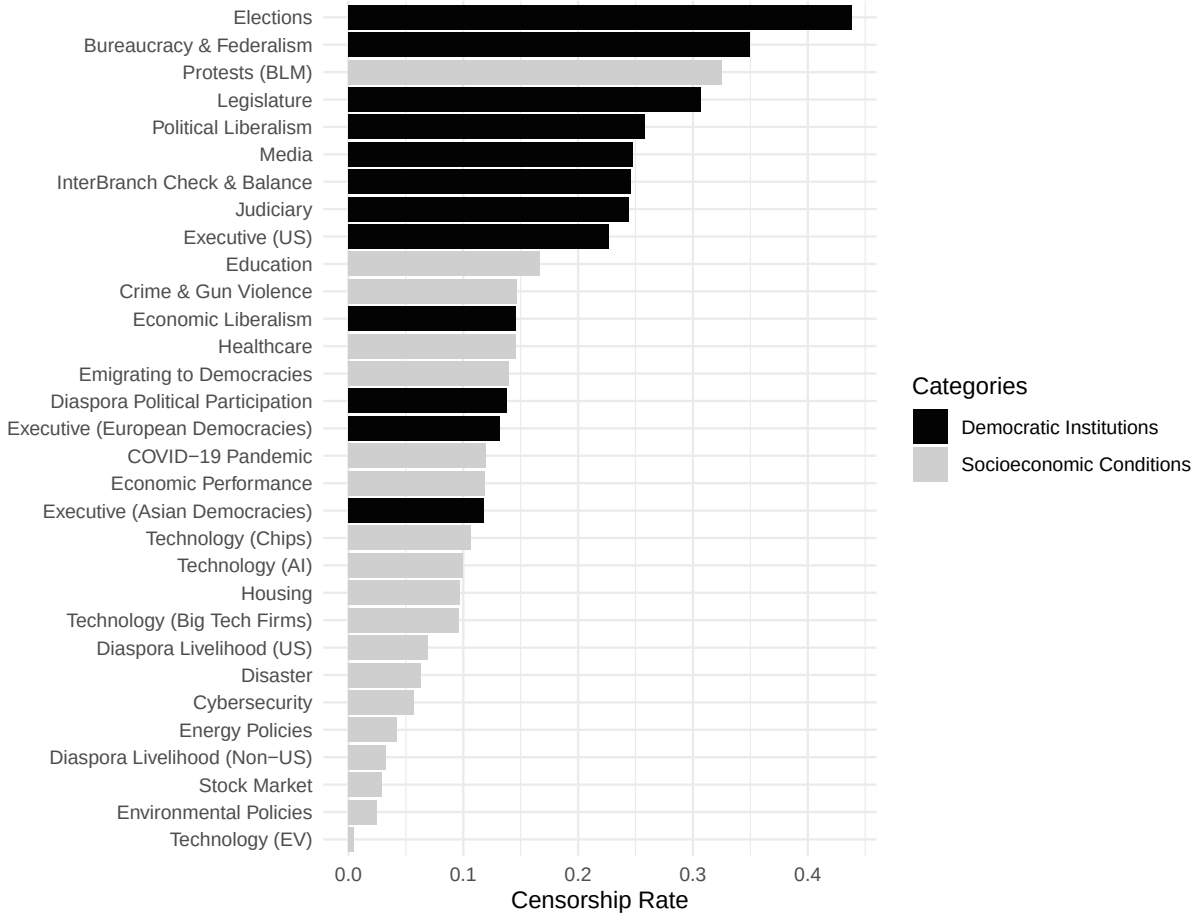


Figure C.2: Probability of Censorship by Specific Topic Categories: Full Censorship Data

Next, Figure C.3 reports the probability of censorship by specific topic categories within democratic institutions and content stances. The most striking difference between the main analysis and the analysis with the full dataset is the *Checks & Balances*. While in the main analysis without China-related articles, positive and negative articles are equally likely to be censored, articles related to checks and balances, China, and are positive toward democracies are significantly more likely to be censored. This suggests that when articles use positive coverage of democratic checks and balances to criticize the Chinese regime, it is much more likely to be censored. Most other topics are consistent with the main analysis.

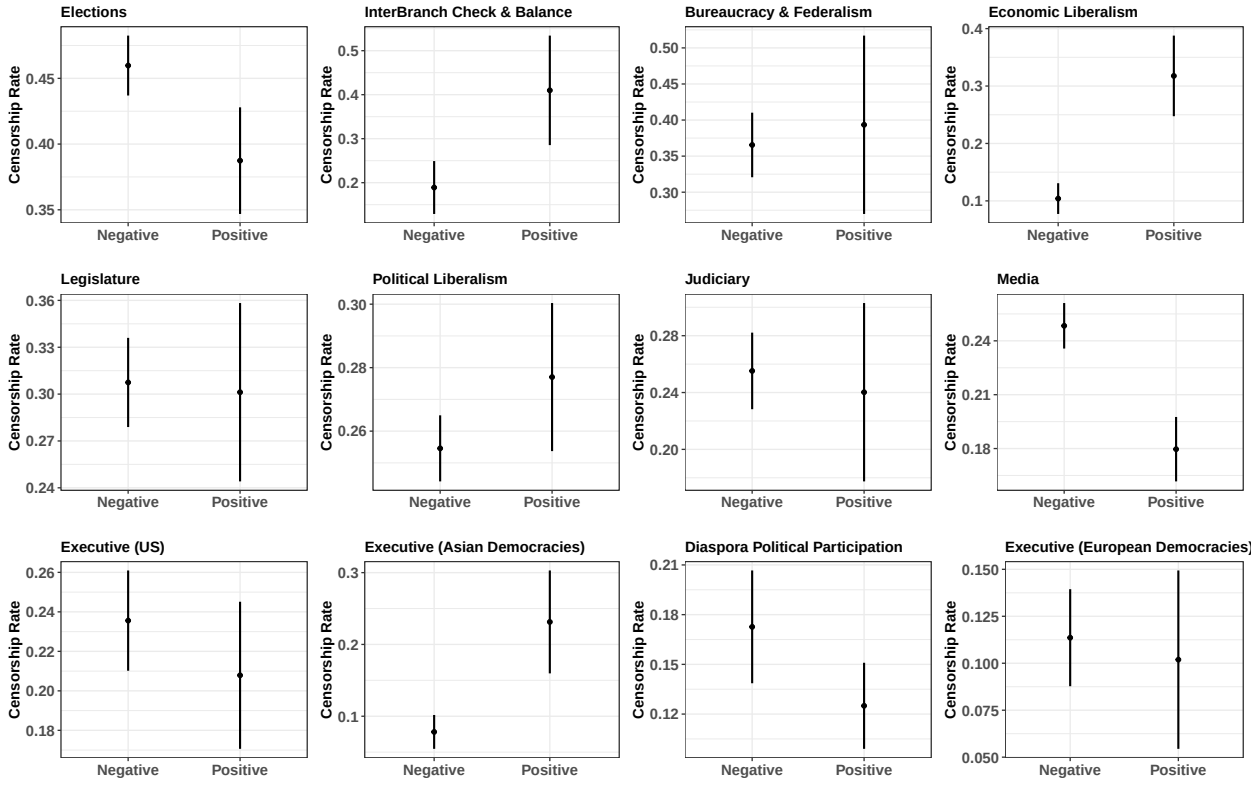


Figure C.3: Probability of Censorship by Specific Topics in Democratic Institutions and Stance: Full Censorship Data

Finally, Figure C.4 demonstrates that the regression analysis results are generally consistent when using the full censorship data.

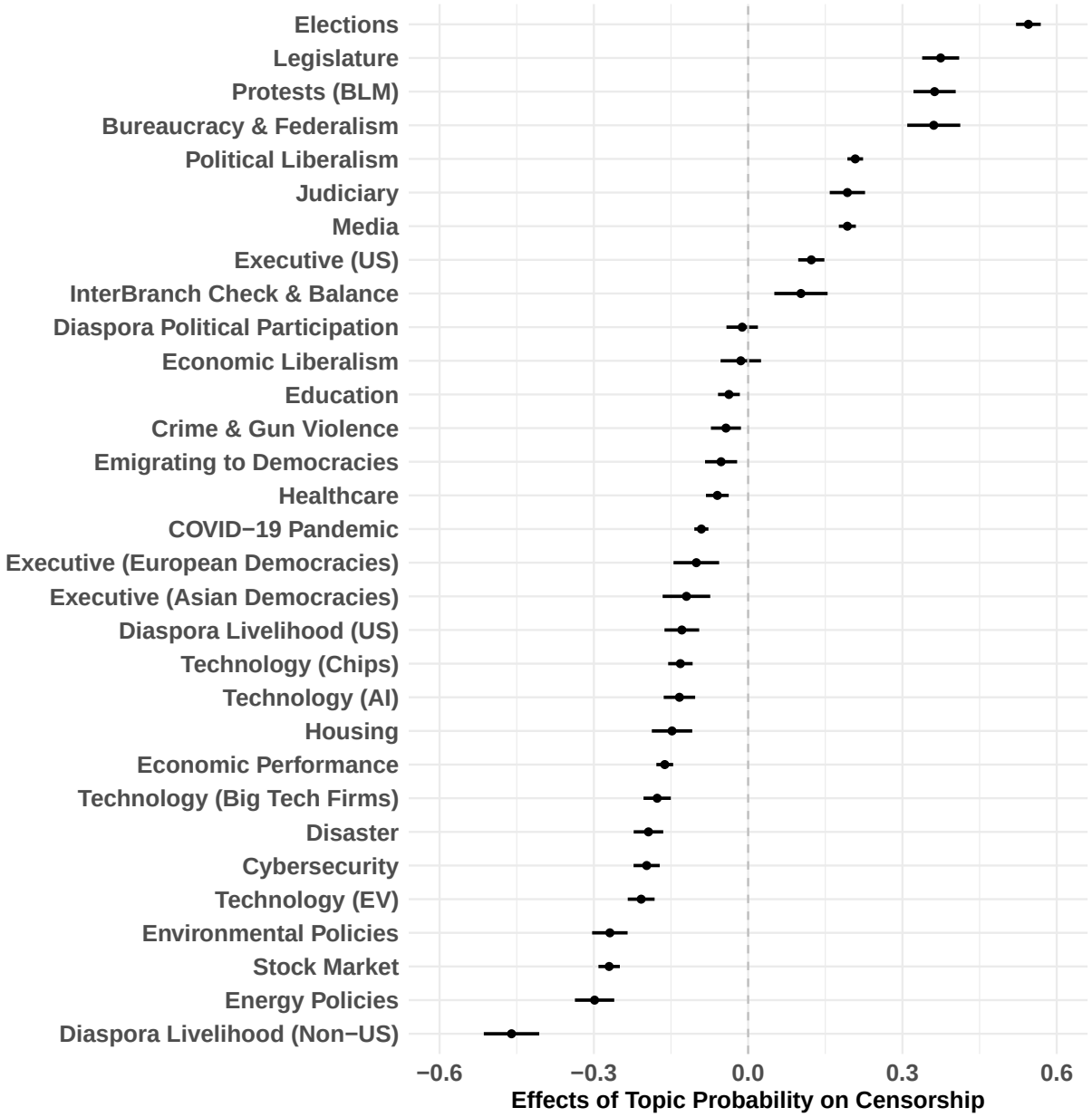


Figure C.4: Effects of Specific Topic Probability on Censorship Incidence: Full Censorship Data

C.3 Alternative Modeling Strategy

In the main analysis, I use OLS models for more straightforward interpretation of the results. Figure C.5 shows the results using binomial logistic regression models. Similar to the main analysis, I control for content stance and the major democratic countries mentioned in the articles. The patterns are generally consistent with the main analysis, as well as the results in Figure C.4.

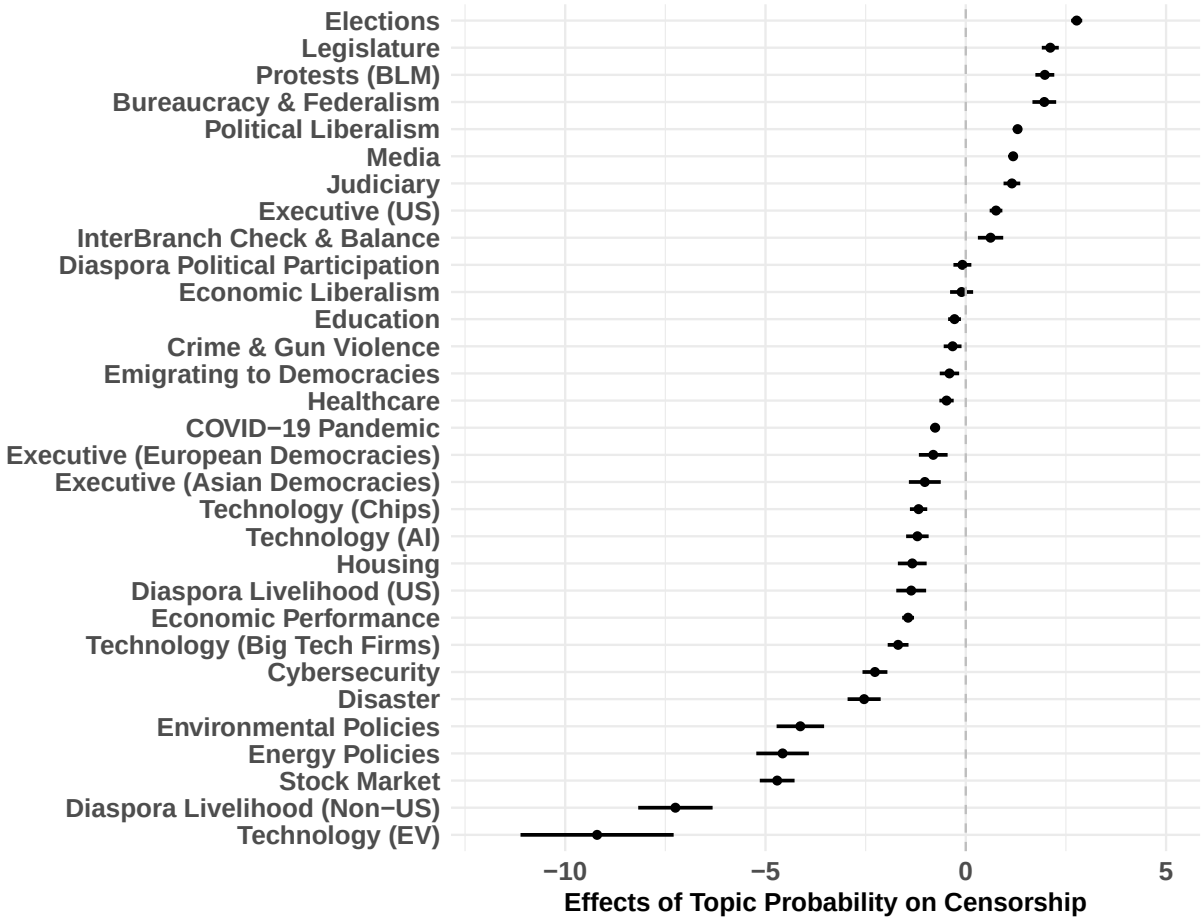


Figure C.5: Effects of Specific Topic Probability on Censorship Incidence: Full Censorship Data with Binomial Logisic Models