Auditing, Understanding, and Improving Recommender Systems on Social Platforms

By

MUHAMMAD HAROON

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

COMPUTER SCIENCE

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Professor Magdalena Wojcieszak, Chair

Professor Hao-Chuan Wang

Professor Anshuman Chhabra

Committee in Charge

2025

© Muhammad Haroon, 2025. All rights reserved.

To my wife, Annum, and to my parents, Naveed and Rehana.

Contents

List of Figures v

List of Tables viii

Abstract x

Acknowledgments xi

Chapter 1. Introduction 1

1.1. Motivation 1

1.2. Auditing Recommender Systems 2

1.3. Disentangling Algorithms and Users 3

1.4. From Diagnosis to Intervention 4

1.5. Contributions 5

Chapter 2. Related Works 7

2.1. Personalization, engagement optimization, and ranking 7

2.2. Auditing social platforms 8

2.3. User agency and pathway-resolved consumption 11

2.4. Interest bias versus ideological bias 12

2.5. Top-down versus bottom-up interventions 14

Chapter 3. A Framework for Auditing Social Platforms 16

3.1. Introduction 16

3.2. Methodology 17

3.3. Data and Measures 22

3.4. Results 22

3.5. Discussion and Implications 29

Chapter 4. Disentangling Users and Algorithms 34
4.1. Introduction 34
4.2. Methodology 36
4.3. Data and Measures 40
4.4. Results 41
4.5. Discussion 48
Chapter 5. Interventions on Social Platforms 51
5.1. Introduction 51
5.2. Background Interventions to Manipulate Recommendations 52
5.3. Nudging Interventions to Promote Healthy Information Diets 59
Chapter 6. Conclusion 75
6.1. Summary of Findings 75
6.2. Implications 76
6.3. Future Directions 77
6.4. Closing Remarks 78
Bibliography 80

List of Figures

3.1 Overview of our sock puppet audit of YouTube's recommendations. 18

3.2 On the left, Spearman's rank correlation (ρ) of 286 video articles from AdFontes and their estimated slant using the Twitter approach. On the right, Spearman's rank correlation (ρ) of 85 channel-level scores from AdFontes and their channel-level slant using the Twitter approach. 19

3.3 Ideological congeniality between slant of training videos and the slant of homepage recommendations. The higher the slope of the red line, the more congenial the recommendations to the training. 23

3.4 Probability (cumulative distribution) of encountering a video of certain slant in the top-n homepage recommendations. For the *very-right* sock puppet, there's a 65.5% chance that the first homepage recommendation is congenial (i.e., has a slant $\geq +0.6$). 24

3.5 The distribution of ideological categories in the top-n homepage recommendations. For the *very-left* sock puppet, the top-1 recommendation is 58% likely to be congenial. 25

3.6 Exposure to ideologically *very-left* and *very-right* videos in the up-next recommendations. The colored lines correspond to the ideology of the sock puppet. The x-axis corresponds to the depth d and the y-axis to the exposure E for the *very-left* and *very-right* ideologies. For example, in the right-hand figure at depth 1, the *very-right* sock puppet is exposed to 1.7× more *very-right* videos. 26

3.7 The mean slant of the ideologically congenial videos watched by each ideology category in the up-next recommendations. For example, the top red line corresponds to the mean slant of the *very-right* videos watched by the *very-right* sock puppet. The dotted and dashed lines correspond to 1-standard deviation and mean slant at depth 1 of the trail respectively. The mean and standard deviation become progressively extreme for the *very-left* and *very-right* sock puppets. 28

3.8 Percentages of problematic video recommendations across a) all recommendations and b) across all sock puppets. The figure on the left shows the percentage of problematic recommendations over all recommendations at each depth. The figure on the right shows the cumulative percentage of sock puppets that encountered one or more problematic recommendations up to that depth. 30

4.1 Distributions of the estimated slant of political videos consumed by participants through each pathway broken down by participant ideology. The annotations in the middle correspond to the mean estimated slant of the distribution. From left-to-right, we see the means move from left-leaning to right-leaning values, indicating that the political content consumed on YouTube through each of the pathways is congenial to users' political ideology. *HPR: Homepage recommendations, SBR: Sidebar recommendations, SQR: Search queries, EXT: external sources, CHN: Channels browsed.* 42

4.2 The proportion of problematic consumption by participants through each pathway. User pathways (EXT and CHN) lead to higher problematic consumption at 54.9% than algorithmic pathways (HPR, SBR, SQR) at 45.1%. The $\mathcal{N}$ values for *Very-liberal*, *Liberal*, *Moderate*, *Conservative*, and *Very-conservative* are 110, 68, 343, 165, 45, respectively. *HPR: Homepage recommendations, SBR: Sidebar recommendations, SQR: Search queries, EXT: External sources, CHN: Channels browsed.* 44

4.3 Histogram (blue bars) showing raw counts and CDF (red line) showing the distribution of where users clicked in the homepage and sidebar recommendation rankings. We see that user clicks are concentrated in higher ranked content for both homepage and sidebar. 47

5.1 The proposed RL based approach. 55

5.2 CDFs showcasing the slant distribution for sock puppets pre-intervention and post-intervention (for different values of m, i.e., the # of intervention videos injected). . . 58

5.3 Example of YouTube Homepage in the User Nudge Treatment Condition 64

5.4 Example of Individual Video Page in the User Nudge Treatment Condition 64

5.5 Overview of the research design. Participants were initially screened for their YouTube browsing habits and political ideology, and then suitable participants were asked to install the ResearchTube browser extension. The extension assigned either treatment or control and then monitored and administered interventions for the next four weeks. 66

5.6 Average daily percentage of News (top row) and Political (bottom row) videos watched by and recommended to participants over total watches and recommendations. W1, W2, and W3 refer to the end of weeks 1, 2, and 3 respectively. The treatment ran between W1 and W3. **(A)** Algorithmic nudge **(B)** User nudge **(C)** Control 68

5.7 Effects of treatments on the consumption of news videos and political videos. Dots represent coefficients and horizontal bars represent 95% confidence intervals. 69

5.8 Political slant distribution of the videos recommended and watched, corresponding to the participants' self-reported political ideology. The green-colored distributions correspond to the slant of recommendations and blue-colored correspond to videos watched. The extreme ends represent the interquartile range and the values in-between are the mean slants of the distributions. We see that for the *very-conservative* and *conservative* participants, the recommendations weren't as partisan as the videos watched, suggesting that participants self-selected into their ideology rather than recommendations guiding them to do so. **(A)** Very-liberal **(B)** Liberal **(C)** Moderate **(D)** Conservative **(E)** Very-conservative 70

List of Tables

3.1 Number of problematic channels in each label 21

3.2 Data collection statistics for 100,000 sock puppets 22

4.1 Dataset Statistics 40

4.2 Proportions of pathways contributing to political consumption across the five user ideologies. We see that `EXT` is the most frequent pathway and `SQR` the most infrequent. *`HPR`: Homepage recommendations, `SBR`: Sidebar recommendations, `SQR`: Search queries, `EXT`: External sources, `CHN`: Channels browsed.* 41

4.3 Mean ideology estimates for homepage and sidebar recommendations and watched content, split by participant political ideology. Data is shown for the top-1 recommendation, top-8 recommendations, and all recommendations across both homepage and sidebar sections. Lower values indicate left-leaning content, while higher values indicate right-leaning content. The table highlights how the recommended content is less congenial than the actual content that participants end up watching but that there's a higher concentration of congenial content in the higher rankings. 44

4.4 KS-test to compare ideological estimates of videos watched with the videos recommended across ideologies. *`HPR`: Homepage recommendations, `SBR`: Sidebar recommendations, `SQR`: Search queries, `EXT`: External sources, `CHN`: Channels browsed.* Asterisks indicate statistical significance *: $p < 0.05$, ** $p < 0.01$. 45

4.5 Percentage of problematic recommendations found in top-1 homepage and sidebar recommendations, top-8 homepage and sidebar recommendations, and all homepage and sidebar recommendations, and in the videos watched by the participants. Lower values in the watched show that participants watched fewer problematic recommendations than what they were recommended. Top-1 and top-8 columns are presented as fractions of number of videos in top-1 and top-8 whereas the other All and Watched columns are as a fraction of homepage, sidebar, and watched combined. 46

5.1 Percentage of *very-left*, *very-right*, and *center* homepage videos encountered pre-intervention and post-intervention for both the left-leaning and right-leaning models. 59

5.2 OLS Regressions Predicting the Effects of Treatments on YouTube Engagement. *** $< .001$ ** $< .01$ * $< .05$. Time spent per active day (minutes) on YouTube (week 1 and weeks 2-3) were $\log_{x+1}$ transformed. 67

Abstract

Recommender systems shape how people encounter news, political information, and problematic content online. Their societal impact depends not only on what algorithms suggest but also on how users navigate and consume content, and on whether interventions can meaningfully reshape information diets. This dissertation brings together large-scale audits, a month-long field study with frequent YouTube users, and bottom-up interventions to advance three contributions.

First, we conduct systematic algorithm audits with trained sock-puppets to measure YouTube's recommendation dynamics at scale, focusing on ideological congeniality, potential radicalization, and exposure to problematic channels. We find that recommendations are often ideologically aligned—especially for right-leaning accounts—while exposure to explicitly problematic channels is limited in volume yet non-trivial in reach. These patterns are consistent with platforms' optimization for engagement and personalization. Second, leveraging granular behavioral data from frequent YouTube users, we disentangle algorithmic from user-driven contributions to actual content consumption. Algorithmic pathways (homepage, sidebar, and search) account for the majority of congenial political consumption, with recommender ranking elevating like-minded content. By contrast, democratically problematic content is disproportionately accessed via user-driven pathways such as channel browsing and off-platform links. Third, we design and evaluate interventions that steer recommender ecosystems without platform cooperation. In a month-long randomized experiment, where we unobtrusively seeded user watch histories histories with verified, ideologically balanced news increased both recommendations to and consumption of news content and enhanced ideological diversity without reducing engagement. Together, these studies provide a unified framework for auditing, attributing, and steering platform ecosystems toward healthier and more diverse information environments.

Acknowledgments

I would like to express my deepest gratitude to my advisor, Professor Magdalena Wojciesak, whose continued guidance, support, and encouragement have been invaluable throughout my PhD journey. Her insights and feedback have significantly shaped my research and academic growth, and have allowed me to analyze problems from a multidisciplinary perspective. Her support and mentorship allowed me to push through the challenges of graduate school, and her faith in my ideas and abilities has been a constant source of motivation that has instilled tremendous confidence. I would also like to thank my committee members, Professors Hao-Chuan Wang and Anshuman Chhabra, for their invaluable feedback and support throughout my dissertation process.

I am grateful to Professor Anshuman Chhabra who has remained an incredible mentor throughout my PhD journey, and am thankful for the opportunity to collaborate with him on multiple projects over the previous years, which has significantly enriched my academic experience. I am highly indebted to Professor Zubair Shafiq for his mentorship and support during the early years of my PhD. He was instrumental in helping me develop my research skills and guiding me through the initial stages of my academic journey, and without his support and encouragement, I would not have been able to reach where I am today. I would also like to thank Professor Fareed Zaffar for his mentorship and guidance when I was a researcher at LUMS. His support and encouragement played a crucial role in my decision to pursue a PhD, and I am grateful for the foundation he helped me build.

I would like to express my sincere thanks to my colleagues at UC Davis, who made my PhD journey enjoyable and fulfilling. In particular, I would like to thank my friends Hadi Askari, Hari Venugopalan, Kartik Patwari, Shaoor Munir, and Yash Vekaria. I am grateful for the stimulating discussions, collaborative projects, and shared experiences we've had that have enriched my academic and personal growth.

I would like to thank my parents, Naveed and Rehana, for their unending love, support, and sacrifices throughout my life. Without their guidance and encouragement, I would not have been able to pursue my dreams or even be the person that I am today. I would like to dedicate this dissertation in memory of my late father, Naveed Anjum, who passed in the final year of my PhD. One of the greatest motivations that pushed me through my PhD was to see how proud you would

be of me the day that I graduate. You were there for me through it all, and losing you shattered me when I was so close to the finish line. In the months since your passing, I've found solace in the words you sent to me when I first started my PhD back in 2021, *"Today you are on victory stand I am proud of you my son"*. While I appreciated these words back then, I've only now come to fully understand their depth and significance. They have helped me come to terms with the fact that you were always immensely proud of me. I love you and miss you dearly, Papa. To my mother, Rehana, and my siblings; Abira, Hassaan, and Waleed, thank you for always being my pillars of strength and for always believing in me.

Above all, I am deeply and forever thankful to my wife, Annum, whose unwavering support, love, and encouragement have been my anchor throughout this journey. Her belief in me has been a constant source of strength and motivation, and I am forever grateful for her presence in my life. When I was overwhelmed with the challenges of graduate school or worried about my father's illness, she was always there to lift me up, remind me of my potential, and help me push through. She has been my rock, my confidante, and my biggest cheerleader. I could not have completed this dissertation without her by my side, and I dedicate this work to her with all my love and gratitude.

CHAPTER 1

Introduction

1.1. Motivation

Recommender systems are central to the modern information ecosystem, shaping what billions of people see online every day. Platforms such as YouTube, Facebook, and TikTok rely on machine learning models that optimize for engagement; surfacing videos, posts, and articles most likely to capture user attention [1, 2, 3, 4]. These personalized recommendations drive between 75% and 95% of consumption on platforms [5, 6] and have been credited with expanding choice and personalizing content [7]. However, they have also raised intense concerns about their social and political consequences. Critics argue that algorithmic curation drives users into "rabbit holes" of increasingly extreme and conspiratorial material, contributing to polarization, misinformation, and democratic erosion [8, 9, 10, 11, 12, 6, 13]. Many observe that social media platforms place users in unique environments characterized by self-curated information flows (*selective exposure* [14]), filtered through one's social network (*homophily* [15]), and reinforced by recommendation algorithms (*filter bubble* [9])—leading to a feedback loop—a potential vicious cycle of reinforcement.

Compared to individual biases and social homophily, recommendation algorithms are the least well understood factor that profoundly influences online exposure and behavior [16]. Recommendation algorithms are designed to optimize user activity and engagement on the platform [7] by personalizing recommendations based on users' past exposures and content viewed or shared by other similar users. Yet empirical findings of the role of recommendation systems in the consumption of harmful content remain mixed: while some studies document evidence of algorithmic amplification of harmful content [17, 18, 19, 12], others highlight the central role of user agency in determining what is ultimately consumed [20, 6, 21].

These concerns are especially salient in the American context where there is a growing gap between the left and the right on key policies [22], hostility between partisans is increasing [23], and

support for political violence and rejection of democratic norms are not uncommon [24]. YouTube, one of the most widely used information platforms in the United States, reaching a majority of adults across demographic and partisan lines [25], has 70% of content consumption through its proprietary recommender system [26]. While it is the most consumed content on the platform [27, 28, 29], the stakes of algorithmic design extend far beyond entertainment: they implicate how citizens access news, form political attitudes, and deliberate about democratic choices. Prior work highlights that exposure to extremist, conspiratorial, or otherwise problematic channels on YouTube is not just an abstract risk but a pressing social problem with potential consequences for trust, polarization, and democratic stability [10, 11, 12]. Under-consumption of public affairs has crucial democratic consequences. Low-information voters are vulnerable to irrelevant cues in the political environment [30], vote against their personal and group interests [31, 32, 33], and are more susceptible to populist, manipulative, and misinformative rhetoric [34].

1.2. Auditing Recommender Systems

A first line of research has sought to audit recommender systems directly, often through the use of sock-puppet accounts that systematically simulate user behavior. These audits have shown that YouTube and similar platforms disproportionately recommend *ideologically congenial content*, particularly to right-leaning accounts, fueling concerns about filter bubbles and echo chambers [10, 20]. Some audits have also found that recommendation trails can include extremist, conspiratorial, or otherwise *problematic channels*, even if such content is relatively rare in aggregate [12, 35]. By systematically documenting algorithmic supply, these studies provide crucial evidence in ongoing debates about whether recommender systems autonomously radicalize users or simply reflect their preferences.

We address these concerns through a controlled, large-scale experimental study. Using more than 100,000 trained sock-puppet accounts, each assigned to one of five ideological profiles, we systematically observe YouTube's homepage and up-next recommendations under fixed watch histories. This design allows the measurement of (a) the degree of ideological congeniality in algorithmic supply, (b) whether recommendation trails become more extreme over time, and (c) whether users are funneled toward channels classified as conspiratorial, radical, or otherwise problematic in prior work. We find

that YouTube's algorithm primarily recommends ideologically congenial videos to partisan users, while still surfacing some diverse and cross-cutting content. Congeniality is strongest in homepage recommendations and increases along watch trails for *very-right* users. Although recommendations become somewhat more extreme deeper in the trail, these shifts are substantively small. A growing share of deeper-trail recommendations comes from channels categorized as problematic (e.g., "IDW", "Alt-right", "Alt-lite", "Conspiracy", "QAnon", "WhiteIdentitarian"), especially for *very-right* users. While such recommendations remain a small proportion overall (at most 2.5%), they are encountered by over 32% of users and up to 40% of *very-right* users.

1.3. Disentangling Algorithms and Users

Audits alone cannot resolve the question of attribution: do algorithms drive harmful exposure, or do users themselves actively seek it out? Recent research emphasizes that both mechanisms matter, but in distinct ways. Analyses of *real-user panel data* capturing the five distinct pathways into consumption—Homepage recommendations, Sidebar recommendations, Search queries, Channels browsed, and External links——show that problematic exposure flows disproportionately through user-driven pathways such as channel browsing and off-platform referrals, rather than from prominent algorithmic recommendations [11, 6]. At the same time, algorithms consistently elevate ideologically aligned content, reinforcing partisan information diets. This duality complicates simple narratives about recommender systems as "radicalization engines," suggesting instead that algorithms and users jointly constitute online information flows. Disentangling these dynamics is critical for understanding both the magnitude of algorithmic effects and the limits of user agency.

Chapter 4 extends beyond audit methods by leveraging a longitudinal dataset of real YouTube users who consented to having their YouTube activity collected by installing a web extension. The collected data captured every pathway a user could take to watching a video including Homepage, Sidebar, Search, Channel Pages, and External Links. This granular data allowed us to disentangle consumption into algorithmic versus user-driven pathways. Our findings show that ideologically congenial content consumption on YouTube is largely driven by algorithmic pathways. Algorithmic recommendations are overwhelmingly congenial for both liberal and conservative users, and users further intensify this pattern by self-selecting even more congenial videos. This occurs because the

system ranks congenial content higher: we observe strong concentrations of like-minded videos in the top-8 homepage recommendations and, to a lesser extent, in top-8 sidebar results. Combined with users' tendency to click higher-ranked content, these ranking dynamics substantially amplify congenial consumption. Additionally, most radical or harmful content consumption is user-driven and not algorithmic in that users opt to watch this problematic content by their own volition.

1.4. From Diagnosis to Intervention

If algorithms and users both shape information diets, then interventions must also operate on both fronts. Top-down approaches focus on platform design, altering the objectives of ranking functions to promote diversity, credibility, or civic value alongside engagement [21, 36, 37]. Such interventions demonstrate that modest algorithmic changes can significantly alter exposure without reducing usage. Bottom-up approaches, by contrast, focus on the demand side: nudging interventions alter the choice architecture of interfaces to encourage reflection, highlight source credibility, or promote exploration of cross-cutting perspectives. Prior research shows that such nudges can meaningfully affect user behavior while preserving autonomy [9, 21, 36]. Together, these interventions offer complementary strategies for recalibrating recommender systems in ways that both enhance accountability and respect user choice.

Building on insights from the audit and user-behavior analyses, Chapter 5 evaluates whether the recommender ecosystem can be steered without platform cooperation using bottom-up approaches. We conduct a month-long randomized field experiment in which a subset of frequent YouTube users received unobtrusive "algorithmic nudges": verified, ideologically balanced news videos inserted into their watch histories at controlled intervals. Evidence from the study shows that algorithmic nudge substantially and durably increases both recommendations and actual consumption of news on YouTube. We observe a reinforcing feedback loop in which recommendations drive news consumption more strongly than consumption drives subsequent recommendations. Additionally, the nudge reduces the ideological congeniality of both recommendations and consumption——an effect strongest among conservatives——and these reductions persist even after the intervention ends. Finally, although the algorithmic nudge successfully increases exposure to diverse and cross-cutting news, this higher consumption does not affect the attitudes or beliefs measured in the study, and these null effects

hold across levels of baseline news exposure and partisan strength. Importantly, the intervention does not reduce time spent or engagement on YouTube, suggesting that such approaches are feasible for platforms concerned with maintaining user activity.

1.5. Contributions

This dissertation makes several contributions to the study of recommender systems, algorithmic accountability, and online information consumption. By combining large-scale audits, real-user analyses, and experimental interventions, it advances both methodological and substantive understanding of how algorithms and users jointly shape exposure to political and problematic content. The key contributions are as follows:

- **Novel auditing framework.** We develop a comprehensive auditing strategy for recommender systems using sock-puppet accounts, enabling systematic measurement of algorithmic supply of content. This approach clarifies when and how recommender systems amplify *ideologically congenial content* and whether they expose users to *problematic channels*, providing an empirical foundation for debates about algorithmic "rabbit holes."
- **Disentangling algorithms and users.** We provide the first large-scale decomposition of exposure into algorithmic versus user-driven pathways. By analyzing *Homepage recommendations*, *Sidebar recommendations*, *Search queries*, *Channel browsing*, and *External referrals*, we demonstrate that recommender systems shape congenial exposure, while problematic exposure is disproportionately driven by user choices. This reconciles contradictory findings in the literature and offers a clearer attribution of responsibility.
- **Experimental interventions.** We design and evaluate user and algorithmic nudges. Through randomized controlled experiments, we show that nudging interventions can modestly increase exposure to diverse and credible sources while preserving autonomy. This provides practical strategies for steering recommender ecosystems toward healthier information environments without requiring platform cooperation.
- **Methodological contributions.** Across the studies, we introduce reusable methodological tools for auditing and intervening in recommender systems, including automated data collection pipelines, measures of ideological slant and problematic exposure, and experimental

infrastructures that simulate real-world platforms while enabling controlled manipulations. These tools extend the methodological repertoire for future research on algorithmic systems.

- **Theoretical and policy implications.** By demonstrating that algorithms and users contribute in distinct ways—algorithms amplify congeniality, users drive problematic exposure—we refine theoretical debates about agency, echo chambers, and algorithmic power. We also provide actionable insights for platform design and governance: interventions must be tailored to the type of harm being addressed, and effective reform requires a portfolio of complementary strategies spanning both algorithmic design and user choice architecture.

Taken together, these contributions advance scholarly understanding of recommender systems and offer practical guidance for building healthier information ecosystems.

CHAPTER 2

Related Works

2.1. Personalization, engagement optimization, and ranking

Personalization refers to the use of recommendation algorithms that learn to suggest content aligned with users' inferred interests or traits, with the primary goal of maximizing user engagement [38]. In practice, these algorithms analyze user behavior and preferences to predict what content is most likely to capture a user's attention and encourage continued platform use.

Personalization and engagement optimization can benefit users by surfacing relevant, interesting, and enjoyable content, thereby enhancing the user experience and increasing satisfaction. These systems are not inherently problematic; when well-designed, they can broaden users' information horizons or introduce them to new genres and communities. Recent evidence suggests that recommendation algorithms on platforms such as TikTok balance between exploration and exploitation, adjusting personalization based on user engagement signals in ways that can be intuitive and explainable rather than uniformly extreme [39]. However, because most users visit platforms primarily for entertainment rather than verified news [40], algorithmic recommendations often reflect and reinforce these preferences, creating closed loops of low-quality, non-news content consumption [41].

Beyond ideological personalization, recommendation systems also exhibit systematic engagement-driven biases that can expose users to low-quality or harmful material. Research documents algorithmic tendencies to amplify sensational, provocative, or negatively valenced content because such material reliably generates engagement [42]. On TikTok, for example, recommender loops can narrow user exposure and reinforce addictive patterns of consumption [43], while on Facebook, clickbait thrives precisely because platforms optimize for attention rather than information quality [44]. More broadly, evidence of algorithmic negativity bias indicates that negative or emotionally arousing content is disproportionately promoted, increasing the visibility of toxic narratives and polarizing discourse [45].

A growing body of research highlights concerns about the societal effects of engagement-optimized personalization. Scholars warn that such systems may contribute to "filter bubbles" or overly homogeneous and congenial content diets [9, 46, 47, 48, 49]. Personalized feeds can narrow exposure, reinforce existing beliefs, limit cross-cutting encounters, and fragment the public sphere [50, 51]. At the same time, ranking optimizations may inadvertently encode or amplify societal biases, leading to inequitable outcomes for marginalized groups [52, 35]. Foundational work in algorithm auditing further demonstrates that platform ranking systems are inherently shaped by opaque design decisions that can systematically advantage certain actors, often without users' awareness [53].

A related set of concerns centers on the possibility that engagement-maximizing algorithms preferentially promote sensational, emotionally charged, or toxic content. Recent work finds that ranking systems may amplify divisive or out-group hostile posts beyond what users would deliberately choose, suggesting that engagement-based optimization can exploit human cognitive biases [54]. Field experiments also show that exposure to toxic content increases user attention and advertising impressions, creating incentives for platforms to retain—rather than suppress—such material [55]. These dynamics contribute to a broader environment in which outrage, negativity, and conflict-ridden content are rewarded by platform algorithms [45]. Even when personalization does not directly induce political radicalization (as evidenced by findings from [56]), algorithms may still shape the distribution and visibility of political actors. For example, research on Twitter's Home timeline shows systematic algorithmic amplification of right-leaning political content compared to left-leaning content across multiple countries [57].

Understanding the benefits and risks of personalization thus requires situating engagement optimization within broader debates about algorithmic governance, democratic resilience, and the societal impact of ranking systems.

2.2. Auditing social platforms

A central question is whether recommender algorithms across major platforms recommend ideologically congenial content, drive users toward increasingly extreme material, and expose them to problematic sources. On YouTube, large-scale sock-puppet and trace-based audits report disproportionate recommendation of ideologically congenial content—particularly for right-leaning

profiles—and mixed evidence regarding exposure to extremist or conspiratorial channels, which appears limited in volume yet non-trivial in reach and can increase with trail depth [20, 58, 12, 10]. Related work shows that YouTube's recommendations amplify ideologically aligned content for partisan accounts, while exposure to problematic channels is modest but persistent [11, 59, 6]. Methodological comparisons highlight that sock-puppet audits can overstate algorithmic influence by omitting user agency, whereas panel data analyses show that external links and subscriptions explain much of the pathway into extremist or conspiratorial content [11, 6, 41]. These findings underscore the need for pathway-aware audits that reconcile divergent conclusions [35].

Complementing these audits, meta-reviews show that between 8 to 10% of algorithmic recommendations pose detectable risks to users [60], including exposure to harmful, discriminatory, or misleading content [61, 35, 62]. These findings underscore that harmful recommendation patterns are not isolated anomalies but systematic properties of engagement-centered ranking.

Recent audits of Twitter's engagement-based ranking system similarly find that algorithmic curation can amplify emotionally charged, out-group hostile, and partisan content beyond what would appear in a reverse-chronological feed. Milli et al. [54] demonstrate that Twitter's Home timeline disproportionately surfaces divisive political material even when users do not explicitly prefer such posts, providing evidence that engagement-optimized ranking can exploit human cognitive biases. Complementary large-scale analyses show systematic amplification of right-leaning political actors across multiple countries [57], suggesting that algorithmic ranking may introduce asymmetries even when underlying follower networks are held constant.

Parallel evidence across other platforms suggests both common mechanisms and platform-specific dynamics. On Facebook, algorithmic filtering reduces cross-cutting exposure and the platform's architecture can blunt the effectiveness of misinformation removal policies [46, 63]. Facebook's ad-delivery optimization also produces skewed outcomes that reflect and reinforce societal biases [64]. On Twitter, observational and experimental work documents echo-chamber dynamics and polarization effects after exposure to opposing views [48, 47], and bot-based audits show partisan asymmetries in amplification [65]. TikTok audits highlight the sensitivity of recommendations to minimal interaction signals and the creation of language and location-specific filter bubbles [66]. Cross-platform studies find that YouTube recommendations spill over through sharing, shaping what

circulates on Twitter and Facebook as well [41]. More broadly, systematic reviews emphasize that auditing methodologies strongly condition conclusions about algorithmic bias, amplification, and harm [35]. Taken together, these audits across platforms demonstrate that recommender systems amplify congenial exposure and can surface problematic sources, but the magnitude and pathways of these effects vary with platform design, user pathways, and measurement methods.

Beyond bot-based audits, recent trace-driven studies of TikTok show that its recommender exhibits a nuanced exploration—exploitation strategy that adjusts personalization based on minimal interaction signals [39]. While personalization is rapid and highly sensitive, these findings suggest that TikTok's algorithm does not uniformly escalate toward more extreme content but instead responds to a combination of watch-time signals and inferred preferences. Such work underscores platform-specific differences in how recommender systems operationalize relevance and personalization, complicating generalized claims about filter bubbles.

Diving into the economy of such problematic content, Beknazar-Yuzbashev et al. used advertiser impression data and found that toxic posts generated substantially more ad exposure on social platforms, creating financial incentives for platforms to leave such content visible rather than demote it [55]. These dynamics help explain why harmful material can persist or resurface despite policy interventions, and highlight the economic dimensions of algorithmic ranking that traditional audits often overlook.

Finally, even in non-social algorithmic systems such as Google search, audits show that users' engagement choices are often more partisan than the content they are exposed to, indicating that selective clicking and navigation—not only algorithmic curation—drive the partisan skew in online information diets [67]. Across platforms, the interplay between algorithmic supply and user demand is more complex than early filter bubble theories suggested. Yet these dynamics also highlight how ranking systems, optimized for engagement rather than user well-being, can steer online experiences in ways that distort public discourse. Consequently, calls have intensified for transparency, accountability, and value-sensitive design in recommender systems [64, 63, 66, 68].

2.3. User agency and pathway-resolved consumption

A growing body of research highlights the importance of distinguishing what the algorithm recommends from what people actually choose to watch. Using pathway-resolved telemetry allows us to disentangle algorithmic from user-driven contributions to actual content consumption. As shown in recent work, algorithmic pathways (homepage, sidebar, search) account for most ideologically congenial consumption, aided by the fact that recommender systems tend to rank ideologically aligned items higher [41, 6]. At the same time, democratically problematic content is disproportionately accessed via user-driven pathways such as channel browsing and external links, rather than through prominent algorithmic recommendations [11, 59, 6]. This evidence suggests that "algorithms help constitute congenial diets, while users themselves drive a disproportionate share of problematic viewing," reconciling divergent findings about whether recommender systems or user agency are more consequential.

Importantly, harmful recommendation patterns often interact with user pathways. Because recommendation systems can infer sensitive traits—such as emotional state [69], substance use [70], or sexual orientation [71]—algorithmic curation may target vulnerable users with content that exacerbates risk, such as addictive material, self-harm related content, or health misinformation [72]. These harms emerge not solely from algorithmic supply but through the recursive interaction of inferred user attributes and subsequent consumption patterns.

Recent large-scale evidence further supports this distinction between algorithmic supply and user choices. Using browsing-history trace data from over 2,200 Americans, Wojcieszak et al. [40] show that 93% of YouTube consumption is non-political, and that partisan congeniality is concentrated almost entirely in the small fraction of news and political content users choose to engage with [40]. Non-political videos exhibit substantial cross-partisan overlap, indicating that ideological segmentation is not a pervasive property of the recommendation system itself but emerges when users selectively seek news or political content.

Related work across platforms reinforces this conclusion. On Facebook, external shares and group dynamics often account for pathways into misinformation exposure more than algorithmic ranking alone [63, 49]. On Twitter, selective following and partisan retweet networks amplify congenial exposure in ways not solely attributable to recommender design [48, 47]. TikTok audits likewise show

that minimal user interaction can rapidly shape For-You feeds toward niche or problematic clusters, emphasizing the interaction of user agency with algorithmic affordances [66]. Across these settings, pathway-aware measurement highlights how platform design and user choice jointly determine actual consumption patterns, and why effective audits and interventions must account for both.

Cross-platform studies further emphasize how user behaviors and platform affordances jointly determine political information flows. On TikTok, political communication is highly participatory and driven by user-initiated interactions, where duet chains and response trees create complex bottom-up diffusion patterns independent of feed ranking [73]. Republicans generate more political content whereas Democrats engage more cross-partisan, illustrating how user communities and not algorithmic targeting structure political pathways. Relatedly, research on Facebook finds that the spread of misinformation is overwhelmingly explained by who chooses to share it rather than by exposure alone: Over 90% of users shared no fake-news URLs during the 2016 election, and sharing was strongly predicted by age and political predispositions rather than recommendation dynamics [74]. Together, these studies underscore that problematic or partisan trajectories often originate from user decision-making, not from algorithmic determinism.

Finally, experimental evidence shows that even substantial changes to algorithmic pathways do not necessarily reshape users' political views. In a large-scale field experiment during the 2020 U.S. election, removing all reshared content from Facebook feeds dramatically reduced political news exposure—particularly from untrustworthy sources—yet produced no detectable changes in political attitudes or polarization [75]. This suggests that while engagement-optimized pathways influence what content surfaces, attitudinal outcomes remain primarily shaped by users' prior beliefs and choices, further highlighting the necessity of pathway-resolved, agency-sensitive approaches.

2.4. Interest bias versus ideological bias

Much debate has centered on whether recommender systems exhibit ideological bias by preferentially recommending like-minded political content or amplifying extremity. Yet a growing body of work emphasizes the importance of "interest bias," whereby platforms optimize for engagement and thus nudge users toward entertainment and away from verified news even absent partisan cues. As shown in recent audits, recommender systems disproportionately surface entertainment content

compared to news, and in doing so reduce exposure to high-quality political information [41, 6]. In practice, "recommendations exert stronger downstream effects on exposure than exposure does on subsequent recommendations," which highlights the leverage of algorithms in shaping information diets [41].

Complementary evidence from Facebook reinforces the centrality of interest bias over ideological bias. Large-scale behavioral trace studies show that once hard-news domains are filtered out, the overwhelming majority of content users encounter consists of soft, entertainment-oriented, or sensational material rather than political information [74]. This skew toward soft content arises both from user preferences and from ranking systems that elevate engagement-maximizing posts, further reducing incidental exposure to verified news. These patterns suggest that the primary structural bias of recommender systems is not ideological curation but the systematic prioritization of attention-grabbing, low-quality material.

Recent field experiments demonstrate that nudging the algorithm to increase the supply of verified, ideologically balanced news can durably increase both recommendations to—and consumption of—news without reducing engagement [21, 37, 36]. This aligns with long-standing concerns about declining incidental exposure to news on social platforms and the civic risks of news avoidance [9, 49]. Cross-platform evidence also shows that engagement-driven ranking systems contribute to skewed incentives and the crowding out of public-affairs content, with similar patterns documented on Facebook and Twitter [46, 47, 63].

Recent nudging experiments also underscore how interest bias is encoded into algorithmic classifications themselves. Platforms often label users as "politically interested" or "not politically interested" based on prior engagement, which then shapes whether news appears in their feeds at all [76]. Interventions that encourage users to follow news accounts can expand their information inventories, increase downstream recommendations to public-affairs content, and sustainably improve news consumption over time. These findings reinforce the notion that recommender systems suppress news exposure not because of ideological filtering but because low prior engagement leads algorithms to infer disinterest.

Together, this literature suggests that beyond ideological congeniality, recommender systems generate structural biases against news, prioritizng entertaining, sensational, or affect-laden content.

Addressing interest bias through algorithmic nudges, supply-side interventions, or preference-aware re-ranking thus represents a key direction for ensuring healthier and more diverse information diets.

2.5. Top-down versus bottom-up interventions

Interventions to address harmful or biased recommendation outcomes can be broadly categorized into top-down approaches controlled by the platform and bottom-up approaches deployable by users or third parties. Top-down strategies include changes to ranking objectives, moderation pipelines, or the demotion of borderline content. While these efforts can reduce exposure to harmful material, they are often opaque, inconsistently applied, and limited by platform incentives [10, 12, 35]. Evidence from Facebook, Twitter, and YouTube shows that even when moderation policies are applied, algorithmic amplification and engagement optimization can blunt their effectiveness [63, 64, 49]. Another major challenge for top-down mitigation is that harmful or low-quality content can be algorithmically attractive: negative, sensational, or provocative material systematically generates higher engagement [42, 45]. As a result, moderation efforts may be counteracted by ranking objectives, particularly when the algorithm detects user traits associated with vulnerability and serves content that reinforces risky behaviors [69, 70, 71]. On Facebook, analyses of fact checker verified misinformation show that even when low-quality or false domains are removed or down-ranked, the platform's soft-news and entertainment-heavy baseline still dominates user exposure [74]. Similarly, on Twitter, algorithmic amplification favors engaging partisan content over informational quality, and does not systematically demote extremist voices relative to mainstream ones [57]. These patterns underscore that platform-level content policies can only weakly counteract the structural incentives of engagement-maximizing ranking systems.

In contrast, recent work highlights bottom-up interventions that can be implemented without platform cooperation. One family of approaches uses algorithmic nudges, where accounts unobtrusively insert verified, ideologically balanced news videos into their watch histories. These nudges significantly increase both recommendations to—and consumption of—news, improve ideological diversity, and do so without reducing engagement [41, 6, 21, 37, 36]. Another family involves preference-aware re-ranking, where recommendations are adjusted to down-weight harmful or low-quality items while preserving topical relevance [77]. Experiments with large language models

show promise in adaptively filtering harmful content and generalizing to emerging harm categories [66, 68, 77]. Finally, reinforcement-learning-based obfuscation strategies demonstrate that carefully chosen "counter-watches" can steer recommendations toward more diverse and less harmful content over time [58]. Additional bottom-up harm-mitigation tools focus on user empowerment and transparent content control. Tools like Gobo allow users to aggregate and customize feeds across platforms [78], while rotating time or site blockers support healthier engagement choices [79]. Machine-learning based classifiers for detecting misinformation, hate speech, and clickbait [80, 81, 82, 83, 84] and browser extensions that automatically hide toxic content [85] offer further mechanisms for mitigating exposure without requiring platform buy-in.

Together, this literature demonstrates that while top-down interventions remain important, bottom-up strategies provide practical and transparent levers for steering recommender systems toward healthier information environments. They are particularly valuable for researchers, journalists, and civil society groups auditing platforms, since they can be deployed independently of platform cooperation and evaluated transparently in real-world settings.

CHAPTER 3

A Framework for Auditing Social Platforms

3.1. Introduction

Recommender systems on social platforms have attracted intense scrutiny because of their potential role in shaping information consumption, political exposure, and democratic outcomes. YouTube, in particular, has been the subject of much debate by scholars, media, and regulators. It is the most popular platform, used by 81% of the U.S. population and with a steadily growing user base [25]. Over 70% of content watched on YouTube is recommended by its algorithm [26], which is proprietary and opaque to users and regulators. There are concerns that the algorithm exposes users to divisive, conspiratorial, and otherwise problematic content [86]. Accordingly, YouTube has been described as a "long-term addiction machine" [87], accused of putting users in "filter bubles" and "rabbit holes" and claimed to be "one of the most powerful radicalizing instruments of the 21st century" [8]. Yet the empirical evidence is mixed: some studies conclude that recommendations amplify problematic exposure, while others find more limited effects [10, 11, 12, 35].

Despite the immense potential impact of YouTube's recommendation algorithm, evidence to support these claims is inconclusive. First, does the algorithm indeed promote exposure to ideologically congenial videos? Second, does the algorithm drive users to increasingly extreme and radical political content? Whereas some studies report that YouTube tends to recommend like-minded and problematic videos [17, 18, 12, 19], others provide evidence to the contrary [10, 88]. For instance, Papadamou et al. [19] find that users have a 6.3% chance of encountering problematic recommendations when starting from moderate content. Similarly, Ribeiro et al. [10] show that many users migrate from ideologically moderate to extreme channels. In contrast, examining online behavioral traces of YouTube users, Hosseinmardi et al. [11] find little evidence that YouTube's recommendation algorithm is driving attention to radical content. These conflicting conclusions are due to subtle but crucial differences in the methodologies. Some work relies on active measurements

using untrained sock puppets (i.e., without any watch history) [10, 88], and thus cannot capture recommendation processes among actual users. In turn, the studies that rely on real user watch activity cannot tease apart the role of algorithmic recommendations from the actions of the user [10, 11, 59]. More recent works on trained sock puppets find moderate filter-bubbles [13].

We design a sock puppet based auditing methodology that allows us to systematically and at scale isolate the influence of the algorithm in recommendations to ideologically congenial and increasingly extreme and problematic content. Trained sock puppets achieve a middle ground between untrained sock puppets and real users and can reconcile conflicting evidence in prior literature. We trained 100,000 sock puppets, watching a total of 9,930,110 YouTube videos from 120,073 channels, to reflect different ideology categories from *very-left* to *very-right*.

Our large-scale audit examines three key questions: Are recommended videos congenial with users' ideology, especially deeper in the watch trail? Are recommendations deeper in the trail progressively more extreme, based on our classification of the ideological slant of each video? And, Do recommendations deeper in the trail direct users to conspiratorial, misinformative, or radical contents, based on extensive validated lists of problematic YouTube channels [88, 10, 11]?

We find that YouTube's algorithm does recommend primarily congenial videos to its partisan users without, however, shielding them away from diverse and ideologically cross-cutting content. This congeniality is most pronounced in homepage recommendations, and increases deeper in the watch trail for the *very-right* users. Although the videos recommended deeper in the watch trial do become significantly more extreme, these increases are substantively small. We also show that a growing proportion of recommendations deeper in the trail comes from YouTube channels categorized as problematic (e.g., "IDW", "Alt-right", "Alt-lite", "Conspiracy", "QAnon", "WhiteIdentitarian", etc.), and this increase is most pronounced among the *very-right* users (hence, less unlikely for the average YouTube user). Although the proportion of these problematic recommendations is rather low (max. 2.5%), they are still encountered by over 32% of users and up to 40% in the case of *very-right* users.

3.2. Methodology

In order to measure the extent to which YouTube's recommendations are ideologically congenial with regard to the user, we conduct a systematic audit of the platform using "sock puppets": automated

browser instances that mimic YouTube users by watching videos and gathering recommendations. Figure 3.1 provides an overview of the audit which consists of three main phases: assigning ideology, training, and then testing sock puppets. The sock puppets are trained by watching videos of a particular political ideology and then their personalized recommendations are tested in a controlled manner. We test two types of YouTube recommendations: (1) those on the homepage; and (2) up-next recommendation trails starting from a seed video.

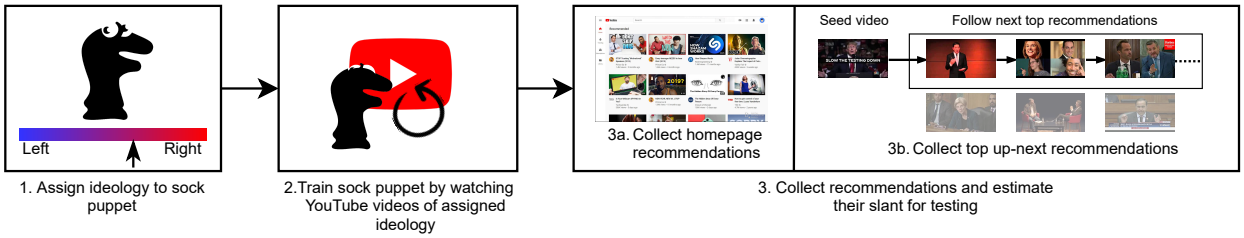


FIGURE 3.1. Overview of our sock puppet audit of YouTube's recommendations.

3.2.1. Estimating Video Ideology. We estimate the ideological slant of a YouTube video by analyzing its audience on Twitter [89]. We emphasize that our approach substantially extends prior work, which analyzes political ideology at a coarse channel-level granularity [10, 88, 11], by estimating slant at the finer granularity of individual videos. Because not all videos posted by a channel reflect that channel's ideology, it is erroneous to assume that each video shares the ideology of a channel (see [90] for evidence on news domains versus individual news articles). In addition, established ideological categorizations exist for only a handful of YouTube channels, and so it is not possible to capture the slant of individual videos from thousands of niche and less popular channels existing on YouTube.

To estimate the video-level slant of any given YouTube video, we collect the tweets mentioning the video using the Twitter API and check if the authors of those tweets follow a set of well-recognized partisan elites, i.e., landmarks, whose political ideology is clearly established as either left (i.e., liberal) or right (i.e., conservative) [91]. To minimize the threat that the estimates are unstable or biased by a particular tweet, we only estimate the slant if the number of landmarks is greater than 12. The landmarks followed by a user provide insight into the user's political ideology, as Twitter users are more likely to follow other users of the same ideology, as consistent with the classic

homophily assumption and the spatial models of ideology that rely on this assumption [89]. For the tweets mentioning a given YouTube video, we count the total number of *very-left* (L_{left}) and *very-right* (L_{right}) landmarks followed by the authors of those tweets and estimate the slant on a scale ranging from -1 (*very-left*) to +1 (*very-right*) as:

$$Slant = \frac{|L_{right}| - |L_{left}|}{|L_{right}| + |L_{left}|} \tag{3.1}$$

In addition to their continuous slant estimate, we also categorize videos into the following slant ranges $[-1, -0.6]$, $[-0.6, -0.2]$, $[-0.2, +0.2]$, $[+0.2, +0.6]$, $[+0.6, +1]$ as *very-left*, *left*, *center*, *right*, and *very-right* respectively. These categories correspond to the sock puppet ideologies we specify during training.

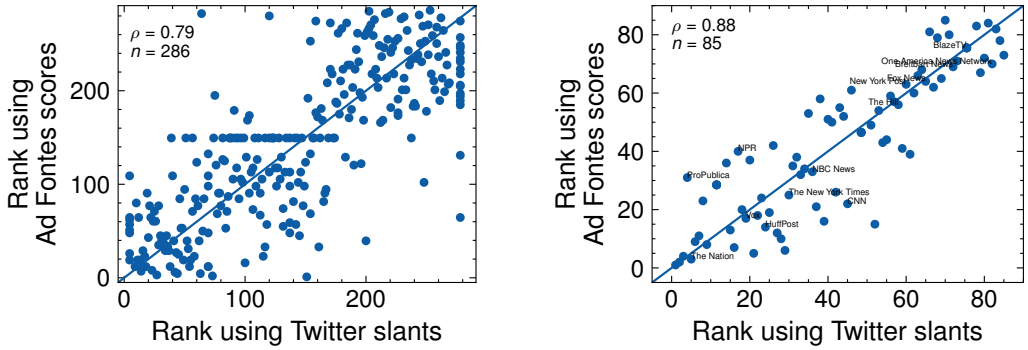


FIGURE 3.2. On the left, Spearman's rank correlation (ρ) of 286 video articles from AdFontes and their estimated slant using the Twitter approach. On the right, Spearman's rank correlation (ρ) of 85 channel-level scores from AdFontes and their channel-level slant using the Twitter approach.

3.2.1.1. *Validation of Slant Estimates.* We assess the validity of our Twitter-based ideology slant estimation by comparing our scores with the Ad Fontes Media Bias Chart [92][1]. Ad Fontes' scores are based on manual labeling of articles, podcasts, and videos, with each individual article and episode rated by at least three human analysts from across the political spectrum (right, left, and center, based on self-reports). The scores are assigned along two dimensions: ideological congeniality and reliability, of which only the former is of relevance to our validation. In addition to article-level ratings, Ad Fontes also assigns a score to news sources based on a weighted average of the pieces of

[1]https://adfontesmedia.com/interactive-media-bias-chart/

content published by that news source. To validate our slant estimation approach, we compare our Twitter-based slant estimates with Ad Fontes scores at both video-level and channel-level.

For validating our slant estimates at the video-level, we sampled 300 videos from Ad Fontes, were able to collect tweets for 286 of them, and then estimated their slant using our Twitter-based approach. Validation at the video-level was then carried out by comparing the ranks of the videos based on our Twitter-based slant estimates with the scores obtained from manual labeling by Ad Fontes. Ideally, when ordered from *very-left* to *very-right* ideologies, the rank of the videos in Ad Fontes labeling should be the same as the rank of the videos in ours. The ranks obtained by both methods for the sampled channels can be seen in Figure 3.2. It is evident that the ranks follow closely along the diagonal ($x = y$), indicating a high correlation between the scores computed by the two approaches. We quantify this correlation by computing Spearman's rank correlation metric [93], which at $\rho = 0.79$ (p-value < 0.05) indicates that the ranks of the two lists ordered left-to-right strongly co-vary. This shows that our automated Twitter-based approach is highly correlated with an expert-based manual coding approach at the video-level.

To validate our slant estimates at the channel-level, we sampled 100 outlets from Ad Fontes, identified their corresponding YouTube channels, and computed a channel-level slant estimate by averaging the slant estimates from the most popular videos on that channel. The ranks obtained by both methods for the sampled channels can be seen in Figure 3.2. We again see that the ranks follow closely along the diagonal and that Spearman's rank correlation metric [93], which at $\rho = 0.88$ (p-value < 0.05) indicates that the ranks of the two lists ordered left-to-right strongly co-vary. This further shows that our Twitter-based slant estimation approach very well aligns with an expert-based manual coding approach at the channel-level as well.

3.2.2. Problematic Channels. For the analysis involving problematic channels, we relied on the manual channel classifications by Ribeiro et al. [10] and Ledwich et al. [88], in similar fashion to Hosseinmardi et al [11]. More specifically, we considered a video to be problematic if it belonged to a channel that was classified by Ribeiro et al. as either *intellectual dark web* (IDW), *Alt-lite*, or *Alt-right*. In the case of Ledwich et al.'s more granular classifications, we considered a video to be problematic if it belonged to a channel that was classified as one of *AntiSJW*, *Conspiracy*, *WhiteIdentitarian*, *Socialist*, *MRA*, *ReligiousConservative*, or *QAnon*. Thus, this analysis was carried

out at a channel-level classification of either problematic or non-problematic. In total, we considered 4,150 channels as problematic, the breakdown of which can be seen in Table 3.1.

Label	Ribeiro et al. [10]	Ledwich et al. [13]	Number of channels
IDW	✓		91
Alt-lite	✓		113
Alt-right	✓		88
AntiSJW		✓	1,056
Conspiracy		✓	2,413
MRA		✓	111
Socialist		✓	148
QAnon		✓	366
WhiteIdentitarian		✓	70
ReligiousConservative		✓	818

TABLE 3.1. Number of problematic channels in each label

3.2.3. Training Sock Puppets. We train sock puppets in the following five ideologies: *very-left*, *left*, *center*, *right*, and *very-right*. Combined, we train over 100,000 sock puppets and each watches a 100 randomly sampled videos from its assigned ideology for 30 seconds each, in accordance with the insight from [12].[2] The videos in each ideology are collected from known politicians, news outlets, pundits, and journalists collected by McCabe and Green (2021) [94].

- List of 747 politicians from the 116th and 117th members of Congress.
- List of 51 state governors.
- List of 617 political pundits
- List of 788 politicians active during the COVID pandemic collected by Gallagher et al.
- List of 6 recent presidential candidates.
- A list of 77 media outlets.

3.2.4. Testing Sock Puppets. Once a sock puppet has been trained, we first collect the video recommendations on its YouTube homepage. The homepage is expected to contain recommendations of interest to the sock puppet based on their watch history [3].

[2]For ads that appear during watch, we skip them if possible and wait until they are over if not.

[3]We note that these sock puppets are not signed into a specific Google account because that requires phone verification which is difficult to do so at scale for 100,000 sock puppets. However, there is evidence that recommendations are personalized even without signing-in through the use of cookies [95]. Although YouTube also uses Google account activity to personalize recommendations [96], because we are not signed-in, it is reasonable to assume that the YouTube watch history is the sole factor influencing recommendations.

Second, we gather the up-next video recommendations from a YouTube video randomly selected from all the training videos (henceforth, referred to as the seed video) and build a recommendation trail, mimicking YouTube's auto-play functionality. Because the sock puppet follows the recommendations, this trail can show whether recommendations are increasingly more congenial and also whether following the recommendations leads users to more extreme and radical content deeper in the trail.

From the collected recommendations, we test whether (i) the homepage recommendations are ideologically congenial with regard to the user; and also whether following the up-next recommendations (ii) increases exposure to ideologically congenial content; (iii) leads to more ideologically extreme content, and (iv) directs users to a growing number of channels that are conspiratorial, radical, or otherwise problematic.

3.3. Data and Measures

Table 3.2 summarizes the data collected from 100,000 sock puppets. In total, we collected over 15 million video watches from nearly 400,000 unique videos across more than 120,000 unique channels.

	Training	Testing	Total
# of videos watched	9,930,110	5,393,820	15,323,930
# of unique videos	23,735	381,153	399,935
# of channels	1,256	119,811	120,073

TABLE 3.2. Data collection statistics for 100,000 sock puppets

3.4. Results

We start by examining congeniality of recommendations in both the homepage and the up-next recommendations. Specifically, we test if the videos recommended to the sock puppets by YouTube's algorithm are consistent with the sock puppets' ideology as determined by the training (e.g., are (very) left sock puppets recommended (very) left videos and are (very) right sock puppets recommended (very) right videos).

Homepage recommendations. Homepage recommendations are shown to the user when they first visit YouTube and are organized in a grid that spans two rows and typically 4 columns. This implies that the user is likely to see 8 recommendations when they first visit YouTube. We thus

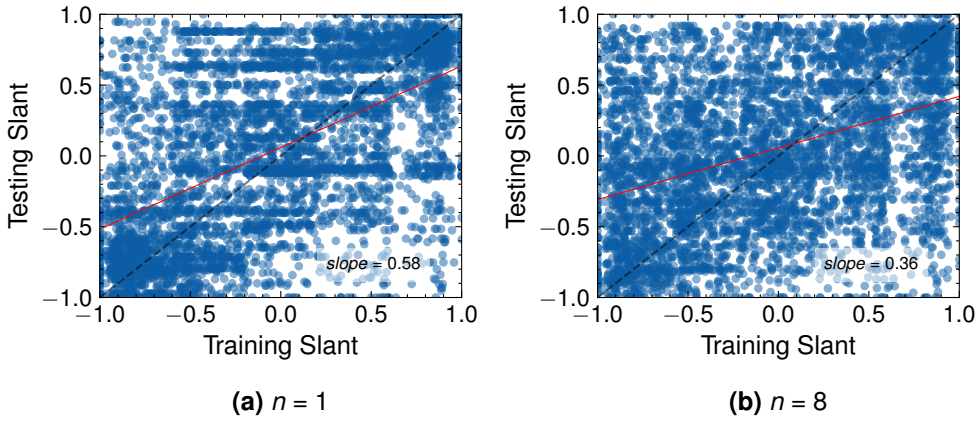


FIGURE 3.3. Ideological congeniality between slant of training videos and the slant of homepage recommendations. The higher the slope of the red line, the more congenial the recommendations to the training.

focus on the first recommendation (top 1, $n = 1$), presumed to be most interesting and important to the user, and all the recommendation a user sees on their homepage without scrolling down (top 8, $n = 8$), analyzing whether these are consistent with the user's ideology.

We first investigate the impact of sock puppet training on the slant of the homepage videos recommended during testing. Figures 3.3a and 3.3b report the congeniality between the slant of individual videos watched in training and videos recommended in testing. We quantify the congeniality in terms of the slope of linear regression (shown by the red line on the figure). A slope of 1.0 (i.e., the diagonal) would indicate that recommendations are perfectly congenial (e.g., 100% *very left/right* videos recommended to a *very left/right* sock puppet). On the other hand, a slope of 0.0 (i.e., a horizontal line in the middle) would indicate that the recommendations are truly diverse (e.g., 50% *left* and 50% *right* videos) or perfectly neutral. (e.g., 100% of videos classified as ideologically moderate). A slope > 0.0 would indicate congeniality in the recommendations and suggest that the training impacts the slant of the videos recommended during testing.

Figures 3.3a and 3.3b show that the slopes are 0.58 and 0.36 for $n = 1$ and $n = 8$, respectively. This suggests that the homepage recommendations are congenial albeit less congenial than the training videos, which—again were exclusively from within one specific ideological category. The fact that YouTube's algorithm does not exclusively recommend 100% congenial videos is not

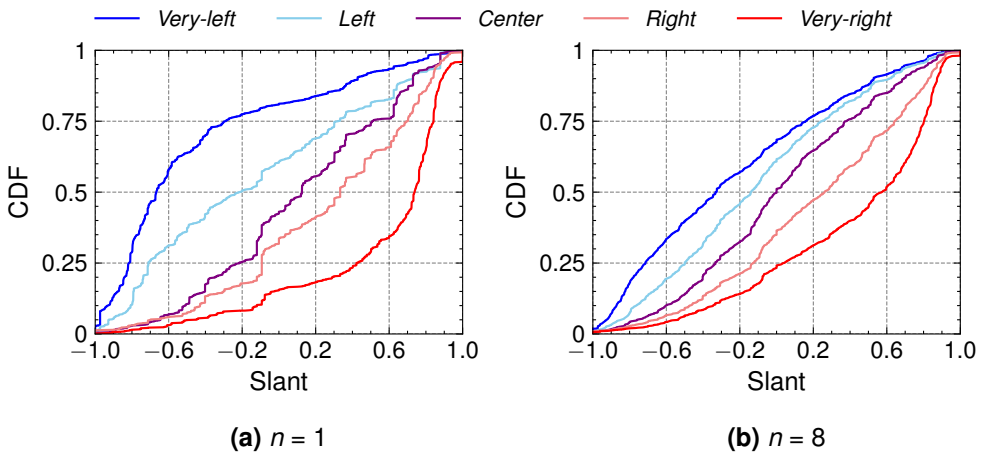
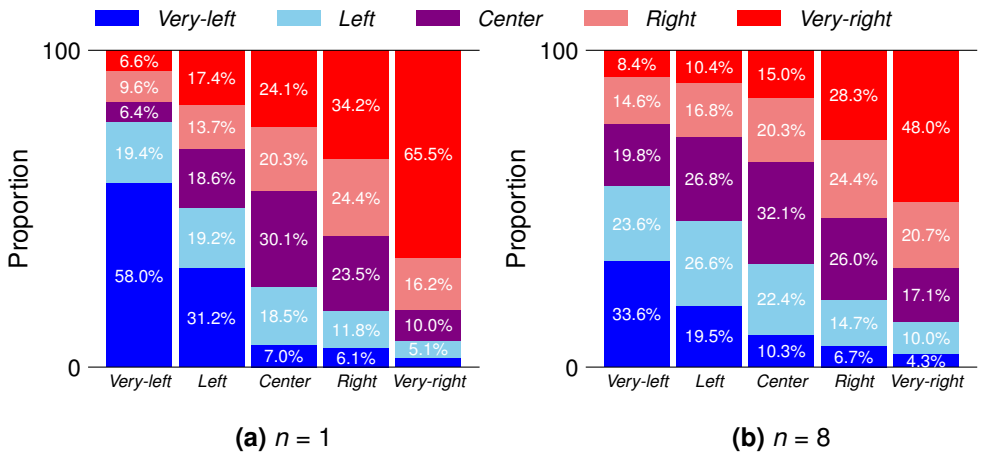


FIGURE 3.4. Probability (cumulative distribution) of encountering a video of certain slant in the top-n homepage recommendations. For the *very-right* sock puppet, there's a 65.5% chance that the first homepage recommendation is congenial (i.e., has a slant $\geq +0.6$).

surprising because the recommender system is designed to recommend "a mixture of personalized recommendations, subscriptions, and the latest news and information" [97, 38, 98]. The exact distribution of recommendations is not disclosed but it is reasonable to assume that the noise introduced by these "trending" or "popular" videos recommended to all users would result in recommendations that are not 100% consistent with the training of our sock puppets.

We next compare the congeniality across sock puppets of different ideologies. Figures 3.4a and 3.4b report the probability that the top-1 and top-8 recommendations on the homepage are congenial. Figure 3.4a shows that for the *very-left* sock puppets, the probability that the first recommendation is also *very-left* is 58.0%. Similarly, the probability that *very-right* sock puppets receive congenial top-1 recommendation is about 65.5%. This figure further shows that the *center* sock puppet sees 30.1% of congenial (i.e., moderate) recommendations but also a high percentage of *right* (20.3%) and *very-right* (24.1%) recommendations.

Going beyond the top-featured recommendation, Figure 3.4b plots the probabilities for the top-8 recommendations. This represents all homepage recommendations a user would see when they first visit YouTube. The probabilities of congenial recommendations are similar albeit slightly lower relative to just the top-1 (33.6% for a *very-left* sock puppet; 48.0% for the *very-right* sock puppet).



(a) $n = 1$ **(b)** $n = 8$

FIGURE 3.5. The distribution of ideological categories in the top-n homepage recommendations. For the *very-left* sock puppet, the top-1 recommendation is 58% likely to be congenial.

The differences between the distributions shown in both figures are statistically significant ($p < 0.05$). Two-sided pairwise Kolmogorov-Smirnov tests between the *center* sock puppet and each of the other sock puppets show that the ideological slant of homepage videos recommended for *very-left*, *left*, *right*, and *very-right* sock puppets is different than that of *center* sock puppets.[4]

These test statistics further indicate that the magnitude of the differences is larger for *very-right* sock puppets as compared to *very-left* sock puppets. Specifically, for $n = 8$ the recommendations for *very-right* sock puppet are significantly more right-leaning compared to the *center* recommendations than those for *very-left* sock puppet.

To offer additional nuance, we plot recommendations to the other slant categories. After all, for instance, recommendations to left-leaning videos are also congenial for *very-left* users. Figures 3.5a and 3.5b show the proportions of the slant categories in top-1 and top-8 homepage recommendations for each of the five sock puppet ideologies.

Extending the above evidence for congeniality, we find that for the *very-left* sock puppets, a total of 77% of the top-1 recommendations are *very-left* and *left*. This proportion is yet higher for the *very-right* sock-puppets, for whom 81.7% of recommendations are *very-right* and *right*. These

[4]We consider the *center* sock puppet as the baseline because of the closeness of its median to zero.

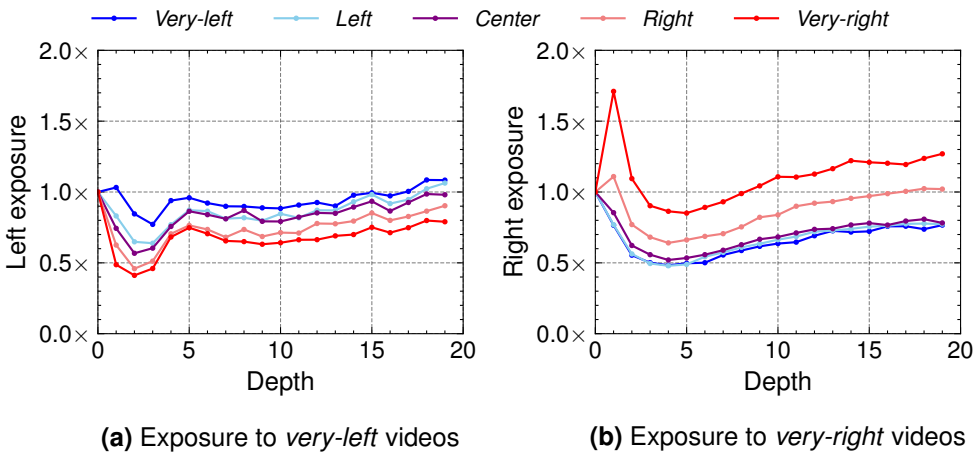


FIGURE 3.6. Exposure to ideologically *very-left* and *very-right* videos in the up-next recommendations. The colored lines correspond to the ideology of the sock puppet. The x-axis corresponds to the depth d and the y-axis to the exposure E for the *very-left* and *very-right* ideologies. For example, in the right-hand figure at depth 1, the *very-right* sock puppet is exposed to 1.7× more *very-right* videos.

proportions slightly decrease when looking at top-8 homepage recommendations, yet still remain the majority (57.2% and 68.7% respectively).

Despite this dominant congeniality, the algorithm also recommends ideologically diverse and even cross-cutting content. For instance, the *very-left* sock puppets encounter *very-right* (6.6%) and right (9.6%) content in the top-1 recommendations (and 8.4% and 14.6% in the top-8 recommendations). These cross-cutting recommendations are less frequent for *very-right* sock-puppets: those have a 3.2% chance of seeing very left and a 5.1% chance of seeing left videos in top-1 recommendations (4.3% and 10%, respectively, in the top-8 recommendations). In general, however, we see an ascending pattern in the percentage of *very-right* and *right* recommendations as we go left-to-right from the *very-left* to the *very-right* sock puppet. A similar pattern—indicative of congeniality—can be observed for *very-left* and *left* recommendations as we go right-to-left.

Up-next recommendations. Do congenial recommendations increase the longer users watch YouTube videos? We now analyze whether following the-up next recommendation trail promotes exposure to a growing number of ideologically congenial content. To measure the degree of this exposure, we define the following exposure metric.

$$E_d^{\text{ideology}} = \frac{\%\text{ of videos}_d^{\text{ideology}}}{\%\text{ of videos}_0^{\text{ideology}}} \tag{3.2}$$

This metric measures the increase in the percentage of videos of a given ideology at depth d over depth 0 (start of the trail). For parsimony, we check for values of ideology $\in \{$*very-left*, *very-right*$\}$ for d up to 20 videos.

Figure 3.6 shows the values of E for both ideology $\in \{$*very-left*, *very-right*$\}$ and as d varies from 0 to 20. There is an increase in the number of *very-right* recommendations as the *very-right* sock puppet follows the trail. The group of *very-right* sock puppets sees the highest values of *very-right* exposures at depth 1 (1.7×), and 1.3× at depth 20. In other words, there is a 37% increase in *very-right* recommendations for the *very-right* sock puppets at the end of the trail, with the overall pattern pointing to a further increase. In contrast, the *very-left* sock puppets initially experience a drop in *very-left* recommendations, a 5% increase at depth 1, and an overall stability in the number of ideologically congenial recommendations deeper in the trail. In sum, following the recommendation trail increases the odds that right-leaning sock puppets will continue to watch ideologically congenial videos, and does not increase recommendations to congenial videos for left-leaning sock puppets.

3.4.1. Radicalization via Recommendations. We now examine whether YouTube's algorithm directs users to increasingly more extreme and politically problematic content. First, we rely on our estimated ideological video slant, which ranges from -1 to +1, to test whether recommendations deeper in the trail become more extreme (i.e., closer to -1 for the *very-left* and closer to +1 for the *very-right* sock puppets).[5] We do this for all five ideology categories. Second, we use the previously compiled lists of 4,150 politically problematic YouTube channels [88, 10, 11]. We consider the following categorizations: "IDW", "Alt-right", "Alt-lite" from the list by Ribeiro et al. [10] and "AntiSJW", "Conspiracy", "MRA", "ReligiousConservative", "Socialist", "QAnon", and "WhiteIdentitarian" from the list by Ledwich and colleagues [88].

Addressing the potential increases in extreme recommendations, Figure 3.7 plots the mean and standard deviations of the distributions at different depths. The mean slant of the videos for the

[5]For parsimony, we consider the distributions of videos corresponding to the sock puppets' ideology. For instance, for the *very-left* sock puppets we look at how the slant distributions for the *very-left* category videos vary as we traverse deeper into the recommendation trail.

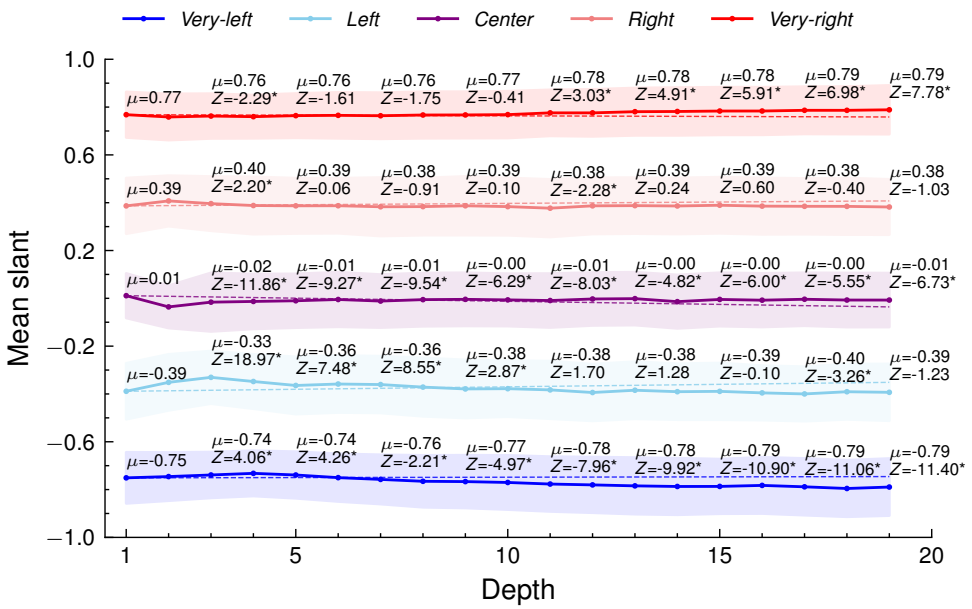


FIGURE 3.7. The mean slant of the ideologically congenial videos watched by each ideology category in the up-next recommendations. For example, the top red line corresponds to the mean slant of the *very-right* videos watched by the *very-right* sock puppet. The dotted and dashed lines correspond to 1-standard deviation and mean slant at depth 1 of the trail respectively. The mean and standard deviation become progressively extreme for the *very-left* and *very-right* sock puppets.

very-left and the *very-right* sock puppets becomes only slightly more extreme as the sock puppets traverse the trail. For the *very-left* sock puppets, it increases from -0.75 to -0.79 and for the *very-right* ones it increases from +0.77 to +0.79. The movement of the standard deviations also suggests that the recommendations become more extreme deeper in the trail for these groups. Although these changes are statistically significant, as reported by the Z-test statistic, they are small in magnitude. For more moderate sock puppets, those in the *left*, *center*, and *right* ideologies, this progression in extremity is not detected.

Addressing the recommendations to the problematic channels, Figure 3.8a plots the percentage of up-next recommendations to all the radical, conspiratorial, or extremist channels from the lists. First, the proportion of problematic channels is rather small overall. Nevertheless, we observe increases in these recommendations deeper in the trail, from 1.2% at depth 1 to 2.5% at depth 10. Although this is not a dramatic increase, the number of recommendations to problematic channels that the sock puppets encounter deeper in the trail is far from trivial in absolute terms. Specifically, by

following the up-next recommendations, the *very-right* sock puppets encounter the greatest number of problematic channels, 22,146, likely due to the majority of these channels being right-leaning, followed by the *right* sock puppets at 19,239. Close behind, the *very-left* sock puppet sees a total of 17,374 recommendations to problematic channels throughout their watch trail, the *left* sock puppet sees 15,132 such recommendations, and the *center* sock puppet sees 16,101[6].

In addition, Figure 3.8b plots the cumulative percentage of sock puppets that encounter problematic recommendations at each depth. We see that this percentage steadily increases at the start before plateauing around depth 10. On average, 36.1% of the sock puppets encounter recommendations to problematic channels in the up-next trail. This number varies between 40% for the *very-right* sock puppet and 32% for the *left* and *center* sock puppets.

Again, although the absolute majority of recommendations are not to problematic content, these patterns suggest that the various radical, conspiratorial, or extremist channels are in fact recommended more frequently the longer the user engages with the platform. Because these channels were *not* used in the sock puppet training, these organic recommendations to problematic content not previously seen are disturbing.

Lastly, we note that the prevalence of problematic recommendations decreases towards the end of the trail, with the decline starting at depth 10 (even though these problematic recommendations are still substantially higher than they were at depth 1). This is likely due to the fact that the channels recommended deeper in the trail are not categorized or included in the original list compiled from [88] and [10].[7]

3.5. Discussion and Implications

Our sock-puppet audit provides a systematic account of how YouTube's recommendation algorithm shapes exposure to political content under controlled viewing histories. Consistent with personalization, homepage and up-next *recommendation trails* surface substantial amounts of *ideologically congenial content*, with the strongest amplification observed for very-right sock puppets.

[6]To generate these numbers, we add the recommendations to the problematic channels between depth 1 and 20 for each category of sock-puppets. The total number of recommendations collected from the *very-left*, *left*, *center*, *right*, and *very-right* sock puppets were 1,650,151, 1,634,174, 1,598,058, 1,560,922, and 1,535,263 respectively.

[7]Specifically, 37.9% of the channels at depth 1 appear in the entire list by [88] and [10], problematic and non-problematic included. At depth 20, this decreases to 21.6%, leading to the detected drop. This likely leads to an underestimation of the actual proportions of problematic channels recommended.

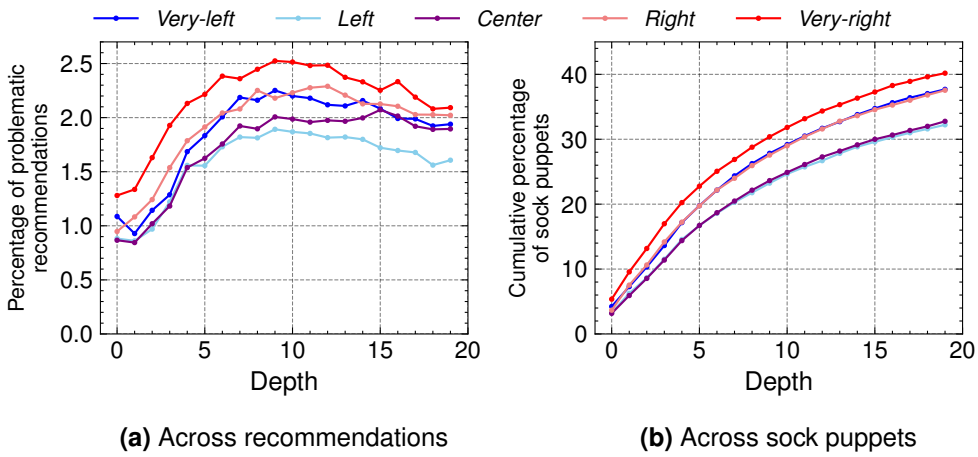


FIGURE 3.8. Percentages of problematic video recommendations across a) all recommendations and b) across all sock puppets. The figure on the left shows the percentage of problematic recommendations over all recommendations at each depth. The figure on the right shows the cumulative percentage of sock puppets that encountered one or more problematic recommendations up to that depth.

At the same time, the algorithm does not exclusively recommend like-minded content: moderate and cross-cutting videos remain present, reflecting YouTube's design to blend personalization with popular or trending items [38, 98]. For instance, we find that the testing videos are less congenial than the exclusively congenial videos on which the sock puppets were trained. Similarly, the users' homepages feature recommendations to moderate and cross-cutting content: even the *very-left* and the *very-right* users may see videos that are diverse or of opposing political ideology (see also [11, 10, 88]). Again, however, ideologically congenial recommendations dominate, especially for the *very-right* users. Right-leaning sock puppets encounter disproportionately more congenial content than left-leaning sock puppets.

Second, it is not the case that congeniality linearly increases as the users traverse the up-next recommendation trail. This finding aligns with a recent audit, which also shows that personalization is more pronounced on the homepage than in up-next recommendations [13]. In our audit, the *very-left* sock puppets do *not* encounter a growing number of *very-left* videos as they follow the up-next recommendations. Nevertheless, for the *very-right* sock puppets, their chances of encountering *very-right* recommendations increase by 37% as these sock puppets continue engaging with YouTube.

This increase is far from trivial and further supports the finding of greater congeniality for this ideological group.

Third, the videos recommended to *very-left* and *very-right* sock puppets do become significantly more extreme the longer these sock puppets engage with YouTube. These shifts, however, are not substantively large and unlikely to be democratically impactful. For instance, the recommendations for the *very-left* sock puppets move from -0.75 to -0.79 between the start of the watch trail and its end (depth 20), and those for the *very-right* sock puppets move from +0.77 to +0.79 on an ideology scale from -1 to +1.

The most consequential risk concerns exposure to *problematic channels*. Although such recommendations are relatively rare in absolute terms at any single depth, they accumulate along trails: more than a third of sock puppets encounter them, and the share rises to roughly two-fifths among very-right sock puppets. Importantly, these channels were not part of training, indicating that the algorithm can *organically* introduce new conspiratorial, extremist, or otherwise problematic sources into watch trajectories [10, 12]. The *very-right* and *right* sock puppets are most likely to encounter them in their watch trail, yet even the moderate sock puppets are recommended a growing number of problematic channels as they follow up-next recommendations. Consistent with past work [59, 11], recommendations to these channels represent a fraction of all the recommendations the sock puppets receive, never surpassing 2.5% on average. Similarly, the increases in these recommendations are not steep, with the maximum increase being 1.3% on average. Furthermore, because 70% of YouTube's content is watched from recommendations [26] without requiring any additional actions or decisions from the users, this pattern of recommendations may activate the cycle of radicalizing exposures. In short, although it is unlikely that an average YouTube user will be steered to more extreme content, the user may encounter videos that promote conspiracy or misinformation, among other problematic messages deeper in the recommendation trail. The finding that this is most likely for far-right users is consistent with Chen et al. [59] who show that "many racially resentful people are not only watching large numbers of videos from alternative or extremist channels, but also are shown recommendations for more such videos when they do so, further increasing exposure to potentially harmful content." Undeniably, the most pernicious potential effects of such recommendations and exposures would emerge among the already predisposed and susceptible users

Fourth, the audit suggests the presence of right-wing bias in YouTube's algorithm. The *very-right* sock puppets encounter more congenial recommendations than any other ideological group, are the only group directed to a growing number of congenial far-right recommendations deeper in the trail, and also are recommended disproportionately more problematic channels than the other groups. We additionally find that the moderate sock puppets encounter more right-leaning than moderate or left-leaning content and that the heterogeneous sock puppets and those trained on one single news source are also directed to more right- than left-leaning content. This finding aligns with recent accusations that some platforms may promote right-leaning content [57] and survey-based evidence that Republicans are twice as likely as Democrats to see right-leaning videos on YouTube [99].

Substantively, these results nuance the "rabbit hole" concern. We observe consistent amplification of *ideologically congenial content* and non-trivial encounters with *problematic channels* along algorithmic trails, especially for right-leaning seeds. Yet we do not find worrisome cascades into extreme content for typical seeds. The recommendation system can present pathways into problematic spaces; whether those pathways translate into sustained viewing depends on behaviors are not modeled. Because sock-puppet audits characterize *algorithmic supply* under controlled histories, they cannot attribute real-world consumption between algorithms and users. Chapter 4 addresses this gap by disentangling algorithmic pathways (homepage, sidebar, search) from user-driven pathways (channel pages, external links) using real-user data to quantify their respective contributions to *ideologically congenial* and *problematic* viewing.

3.5.1. Legal Considerations. YouTube's Terms of Service (TOS) [100] allow the use of the service as long as the user agrees to certain terms. Any violation of the TOS of a website could be considered a violation of the Computer Fraud and Abuse Act (CFAA, 18 U.S. Code § 1030). In other words, violations of TOS such as unauthorized access or exceeding authorized access to a protected system might be penalizable under CFAA. While YouTube is considered a protected system under CFAA, it remains unclear whether our auditing approaches engage in unauthorized access [101]. As note below, overly broad interpretations that attempt to equate TOS violations to crimes under CFAA have been turned down by federal courts in the past.

In terms of the legality of audit by researchers, YouTube's TOS restrict access to the service via "automated means", which means that our sock puppet based auditing system is in violation of

YouTube's TOS. Such violations are subject to penalty under the CFAA because they involve probing a computer system in a misleading manner [102]. To remove overbroad interpretations of the CFAA, the case of Sandvig v. Barr [103] attempted to restrict the use of CFAA to criminally prosecute researchers who violate TOS based on the argument that it violates their "First Amendment right to engage in harmless false speech". In 2020, a federal court ruled that the violation of a website's TOS is *not* grounds for being held criminally liable under the CFAA [104]. This landmark decision provides legitimacy to sock puppet based auditing of algorithmic systems, which is now considered a well-established methodology in the academic research community [10, 12, 19]

Based on this ruling, we argue that the users of our intervention cannot be held liable under the CFAA because they do not exceed in their authorization to access publicly available YouTube.

CHAPTER 4

Disentangling Users and Algorithms

4.1. Introduction

A central debate about YouTube and similar social platforms concerns whether exposure to problematic content is primarily driven by recommender systems or by the active choices of users. On one side, critics argue that YouTube's recommendation algorithm functions as a "radicalization engine," nudging users toward increasingly extreme, conspiratorial, or otherwise problematic material [10, 11, 12]. On the other, scholars emphasize user agency, pointing to evidence that individuals often seek out ideologically congenial or sensationalist content on their own [20, 21]. This debate is crucial for attributing responsibility: if algorithms are the main driver, then reforming recommender design is the obvious lever; if users themselves are primarily responsible, then interventions must focus instead on demand-side behaviors. Furthermore, this understanding is critical to the ongoing policy discussions about the impact of platforms on users.

We address this debate by systematically disentangling the role of algorithms from the role of users in shaping exposure to political and problematic content on YouTube. Combining data about YouTube's algorithmic recommendations (both homepage and sidebar) *and* users' watches to tease apart the role of the algorithms from the role of the users in the consumption of congenial content (i.e., content aligned with users' political ideology) and of democratically problematic YouTube channels (i.e., those promoting conspiracies, extremism, white supremacy, among others). A large sample of 1,866 frequent YouTube users recruited by YouGov installed our web browser extension for four weeks. The extension unobtrusively recorded the participants' interactions with the platform, collecting 2,384,130 videos they were recommended (both homepage and sidebar), 15,887 search queries the users made, and 111,591 videos they actually watched during one month.

Our approach is grounded in the five distinct *pathways* that can lead to content consumption: the homepage, the up-next autoplay trail, sidebar recommendations, direct navigation to a channel

page, and external links from outside platforms. The first three pathways are primarily algorithmic, as they reflect the outputs of YouTube's recommender system [38, 98]. The latter two pathways are user-driven, reflecting deliberate acts of navigation that bypass algorithmic selection. This pathway framework allows us to move beyond aggregate measures of exposure and instead ask where problematic content originates. Do algorithms surface problematic material prominently on the homepage or in the up-next trail? Or do users arrive at such content primarily through channel browsing, or external referrals? Distinguishing between these mechanisms is essential for clarifying whether harmful exposure reflects *algorithmic supply* or *user demand* [35].

Furthermore, we rely on a BERT-based classifier to identify political videos, which were recommended to and watched by users, and on our previously developed approach to at-scale classify the ideological slant of all these videos in our data [40]. This, combined with self-reports of users' ideology, allows us to determine consumption congeniality, i.e., right[left]-leaning users watching right[left]-leaning videos. To determine recommendations to and consumption of problematic content, we identify videos from curated lists of problematic channels (e.g., "Intellectual Dark Web (IDW)," "Alt-right," "Conspiracy," "QAnon," etc.) [88, 10] used in past work [6, 11].

These data together allow us to test whether it is primarily YouTube's algorithms (i.e., homepage and sidebar recommendations, searches) or the users themselves (i.e., channels browsed, browsing on off-platform external sources) that drive the consumption of congenial and problematic content. Furthermore, the data we have on all the videos recommended to each user allow us to show whether—given a set of recommendations—users self-select to more or less congenial and problematic videos than what the algorithms recommend, similar to how Hosseinmardi et al. [11] showed that problematic viewing patterns are more reflective of user preferences. Lastly, we also examine how the ranking of the recommendations influences this potential self-selection.

Specifically, we address the following three research questions. Do users consume (a) ideologically congenial and (b) problematic content on YouTube? If so, is it primarily through algorithmic (i.e., homepage and sidebar recommendations, search queries) or user-driven (i.e., channel browsing, external sources) pathways (RQ1)? From within the platform's algorithmic recommendations, do users self-select to more or less (a) ideologically congenial and (b) problematic content than what

is recommended (RQ2)? Does the ranking of recommendations, contribute to more or less (a) ideologically congenial and (b) problematic content consumption (RQ3)?

We offer two key findings. First, ideologically congenial content consumption on YouTube is largely driven by recommender systems. Not only a higher proportion of congenial consumption comes from the algorithmic pathways (total of 62%; 21.7% from homepage, 35.1% from sidebar recommendations, 5.2% from searches) than from the user-driven pathways (total of 38%; external sources 32.7%, channels 5.3%) but also algorithmic recommendations are overwhelmingly congenial for both liberal and conservative users, although we also show that the users self-select into yet more congenial content than what the algorithms suggest, this is largely because the algorithmic system ranks congenial videos higher. We find a higher concentration of congenial content in the top-8 homepage recommendations and, albeit to a lesser extent, top-8 sidebar recommendations. This, combined with the users' tendency to click on higher-ranked content, results in especially high levels of congenial consumption, suggesting that this feature of the system design contributes to congeniality.

Second, similar to the findings of [6, 11], the majority of problematic consumption is driven by the users themselves (total of 54.9%; external sources 46.1%, channels 8.8%), not by the algorithmic pathways (total of 45.1%; 18.2% by homepage, 23.1% by sidebar recommendations, 3.8% by searches). Only a small percentage of problematic videos can be found on the homepage and sidebar recommendations, further underscoring the role of the users in the consumption of radical, misinformative, conspiratorial, and otherwise problematic YouTube content.

4.2. Methodology

To disentangle the impact of YouTube's recommender system and the users' agency in naturalistic settings, we developed a Chrome extension for YouTube. Participants installed the extension as part of the survey which monitored their YouTube behavior for 4 weeks. The extension collected recommendations, both homepage and sidebar, and also recorded all user searches and video watches throughout that time. The extension did not collect any off-platform (non-YouTube) activity. All the data were stored with a unique user identifier, which was linked to users' survey responses.

We deployed this extension for one month on a sample of frequent YouTube users (i.e., those who visited YouTube more than once a week) recruited by YouGov between November 2022 and January 2023. The 2,142 eligible participants were directed to the Chrome Store, where they were prompted to install the extension for four weeks. We analyze data collected from 1,866 participants after removing participants who had no YouTube activity (no videos watched and no recommendations collected) during the four weeks.

The analyzed sample (N = 1,866) was 51% female, 69% White, and 64% had some college education or a bachelor's degree, with the mean age 46.7 (*median* = 47, *s.d.* = 16.7). Also, 20.6% (n = 385) self-identified as *Very-liberal*, 19% (n = 355) as *Liberal*, 38.9% (n = 725) as *Moderate*, 14.1% (n = 264) as *Conservative*, and 7.3% (n = 137) as *Very-conservative*. We acknowledge that the sample is not representative of the general American population, as no studies using online samples willing to install browser extensions can claim representativeness. Because our objective was to have a sample of active YouTube users, YouGov purposefully recruited only those panelists who reported using YouTube at least once per week. These additional analyses indicate that our sample well reflects the YouTube population and is not too different from the American population, apart from the under-representation of those with low education and over 65 years old, and the over-representation of 18-29 year olds and non-Hispanic Whites (as typical in online studies).

4.2.1. Political Classifier. Given that extant worries about congeniality and radicalization pertain to political consumption and recommendations, we train a BERT-based neural binary classifier that categorizes YouTube videos as related to politics with high accuracy (.93, Precision .92, Recall .91, F1 .915). A dataset of 10,629 Reddit titles was manually labeled by trained coders for whether they were about political issues or not. We conceptualize politics rather broadly as videos related to political figures, political parties, or political events such as elections *as well as* consequential social and public policy issues regarding race, gender, sexuality, ethnicity, and religion [40].

4.2.2. Problematic Channels. We identified recommendations and user consumption of problematic channels from the lists compiled by [10] and [13]. Specifically, [10] categorize channels into "Intellectual Dark Web (IDW)," "Alt-right," and "Alt-lite." and [13] categorized channels into

a list of "Anti-Social Justice Warrior (AntiSJW)," "Conspiracy," "Mens Rights Activist (MRA)," "ReligiousConservative," "Socialist," "QAnon," and "WhiteIdentitarian". In total, 3.78% of political recommendations (14,714 out of 389,074) and 4.88% of all political videos (731 out of 14,985) watched by participants came from one of these problematic channels.

4.2.3. Estimating Political Slant of Videos. To determine whether recommendations and consumption were ideologically congenial, we used the updated approach developed in recent work [40]. We estimate the slant of all political videos in our dataset by analyzing the audience of each video on Twitter and the political elites these audiences follow. We collected the tweets containing a given YouTube video and checked if the authors of those tweets follow any political elites from our a list of 1,604 American politicians, news media, and pundits with their corresponding ideology scores [94]. We collected the followers of these 1,604 elites for a total 16,472,360 unique followers and a total 862,409,644 following relationships. If a video was shared by these 16,472,360 users, the video's slant can be ascertained. The slant score of a video (v) is the weighted average of the ideology of all elites ($\sum_{i=1}^{N_{pol}}$) followed by the users ($\sum_{u=1}^{N_u}$) who tweeted the video(v):

$$v = \frac{\sum_{u=1}^{N_u} \sum_{i=1}^{N_{pol}} Pscore_{i,u}}{N_u N_{pol}} \quad ; \quad where$$

$Pscore(i, u)$ denotes the ideology score of the elite i followed by the user u, N_u denotes the total number of users who tweeted the video, and N_{pol} denotes the total number of elites followed by each user.

4.2.4. Labeling Pathways. There are five pathways to content on YouTube and the data collected by the extension allow the identification of which specific pathway led a user to each video. Although platform recommendations and user content consumption are naturally intertwined and recursively influence each other in nuanced ways These pathways are either algorithmic (i.e., recommendations and searches) or user-driven (e.g., channels browsed, external sources). We identified them by backtracking from the video page and examining the contents and URLs of the webpages the user visited immediately prior to watching a video.

4.2.4.1. *Algorithmic Pathways.* These are pathways to consumption that stem from some form of platform recommendation.

- Homepage recommendation (`HPR`): The homepage is the default YouTube landing page where personalized recommendations are provided to the user. This is an algorithmic pathway as the recommendations are generated by YouTube's algorithms. We determined that a watched video came from the homepage by backtracking from the video page until reaching the YouTube homepage visit within the past five minutes (i.e., when the URL was of the following format `youtube.com/`) and checking if the video was part of the homepage recommendations.
- Sidebar recommendation (`SBR`): A sidebar recommendation is a video that is recommended on a sidebar to the right of the YouTube page while the user is watching another video. Sidebar recommendations are also generated by YouTube's algorithm. We identified that a watched video came from sidebar recommendations by backtracking from the video page until the first page we reached was another video page within the past five minutes, i.e., the URL is of the following format `youtube.com/watch?v=videoId` and seeing if the video was part of the sidebar recommendations.
- Searches (`SQR`): Searching for a video redirects the user to a page where the search results can be seen. The video results are ranked by a recommendation algorithm. We identified videos watched through searches by backtracking from the video page until a search results page was reached within the past five minutes, i.e., the URL was of the following format `youtube.com/search?q=query` and checking if the video was among the search results. We follow past work [105] and classify searches as algorithmic because, similarly to homepage and sidebar recommendations, which are algorithmic but relevant to some user input (i.e., the user's watch history), search results are also algorithmic rankings curated by the platform and relevant to the user's query and personalized to their watch history.

4.2.4.2. *User Pathways.* These are pathways driven by the user's actions.

- Channels (`CHN`): Users can subscribe to YouTube channels and/or watch videos from certain channels by browsing to the channel pages. We identified channels browsed by backtracking from the video until we reached a channel page within the past five minutes, i.e., the URL was of the following format `youtube.com/channel/channelId` and was consistent with the

the video's channel. Because we do not explicitly check whether the user has subscribed to these channels, we refer to this pathway as the channel browsed by the user.

- External (`EXT`): Some video views are driven by non-YouTube factors such as external sources. For example, users may come across a YouTube video while browsing another platform or reading a news website, and decide to click on the embedded YouTube link. External pathways are harder to track as our extension collected data from YouTube, not from other platforms or websites that the participants browsed. We marked all videos that did not fall under any identifiable pathway from above as `EXT`. See [11] for a similar approach.

4.3. Data and Measures

Overall, we collected data from 1,866 participants on 2,384,130 recommended and 111,591 watched videos, as well as information on 3,345 YouTube channels the participants browsed and 15,887 search queries they used. The average viewership per participant was 64 videos with a maximum of 1,154 and a minimum of 1. Table 4.1 provides statistics based on our dataset. We estimated the slant of 315,365 videos, which is 78% of political videos in our data (31.5% more coverage than the aforementioned channel-level lists at 46.5%. In particular, we show that our video-level slant estimation matches with labels from expert human annotators. We also plot the distribution of the slants estimated for the videos from nine news sources, showing that the slant of individual videos may be independent of the channel's.

Number of participants	1,866
Number of watches	111,591
Number of political watches	14,985
Number of recommendations	2,384,130
Number of political recommendations	389,074
Number of channels	3,345
Number of search queries	15,887

Table 4.1. Dataset Statistics

With regards to the political classifier, we identified 404,519 political videos in our dataset (representing 16.2% of the 2,495,721 total videos).

4.4. Results

4.4.1. Consumption Pathways. We first offer aggregate evidence on how users arrived at the political videos they watched on YouTube; whether mostly through recommendations or through searches, channels browsed, or external sources. The dominant pathways to content across all users are external sources `EXT` at 33.2% followed by sidebar recommendations `SBR` at 32.9% and then the homepage recommendations `HPR` at 19.5%. In contrast, the use of YouTube search `SQR` at 7.0% and channels `CHN` at 7.4% was relatively infrequent across all users. Table 4.2 shows the proportion of video watches via each pathway, broken down by participant ideology. Offering initial insight, these trends suggests that algorithmic content consumption (59.4% across `HPR`, `SBR`, and `SQR`) is generally higher than user-driven consumption (40.6% across `EXT`, and `CHN`) with this difference being statistically significant ($Z = 32.52$, $p < 0.01$ for a two proportion Z-test).

	HPR	SBR	SQR	EXT	CHN
Very-liberal	19.4%	33.2%	6.8%	35.6%	5.0%
Liberal	19.0%	37.3%	9.1%	28.9%	5.8%
Moderate	19.5%	31.0%	6.5%	32.9%	10.0%
Conservative	19.1%	30.6%	6.1%	37.4%	6.8%
Very-conservative	23.8%	36.7%	7.2%	26.4%	5.9%
Overall	19.5%	32.9%	7.0%	33.2%	7.4%

TABLE 4.2. Proportions of pathways contributing to political consumption across the five user ideologies. We see that `EXT` is the most frequent pathway and `SQR` the most infrequent. *`HPR`: Homepage recommendations, `SBR`: Sidebar recommendations, `SQR`: Search queries, `EXT`: External sources, `CHN`: Channels browsed.*

4.4.1.1. *Congenial Consumption.* Political content consumption on YouTube is highly congenial [40]. Figure 4.1 shows the mean video slant estimate move from left-to-right in correspondence with liberal to conservative ideologies of the participants. The *Very-liberal* and *Liberal* participants consume significantly more left-leaning videos than the *Moderate* ones, and the *Conservative* and *Very-conservative* participants consume significantly and substantially more right-leaning videos than *Moderate*. The majority of the differences between the slant of videos consumed by *Moderate* participants versus the other groups are statistically significant.

Breaking the overall consumption down by the five pathways, Figure 4.1 also shows that `CHN` pathways are the most congenial, followed by `HPR` and `SBR`. For the *Conservative* and *Very-conservative*

participants, we see more congenial consumption through the user-driven CHN pathway, and a higher mean for EXT pathway for the *Conservative* participants. This suggests that conservative users largely chose to watch ideologically congenial content on YouTube. We do not see clear patterns for *Very-liberal* and *Liberal* participants, for whom the algorithmic and the user-driven pathways drive similar proportions of congenial consumption.

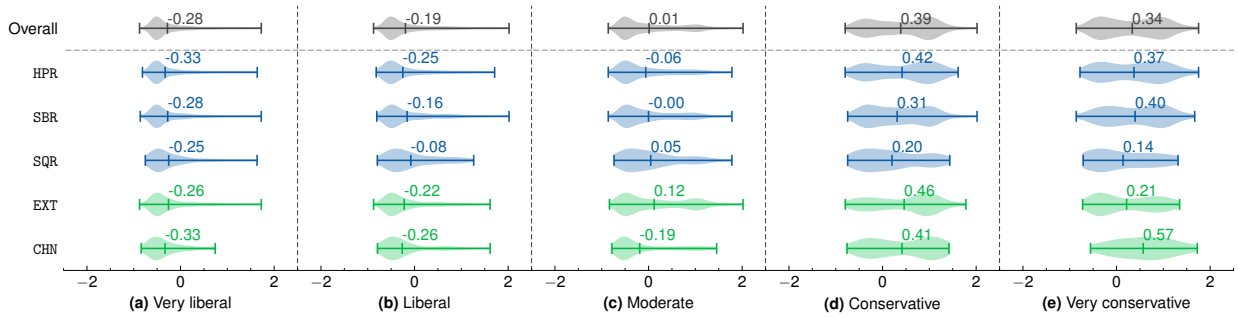


FIGURE 4.1. Distributions of the estimated slant of political videos consumed by participants through each pathway broken down by participant ideology. The annotations in the middle correspond to the mean estimated slant of the distribution. From left-to-right, we see the means move from left-leaning to right-leaning values, indicating that the political content consumed on YouTube through each of the pathways is congenial to users' political ideology.
HPR: Homepage recommendations, SBR: Sidebar recommendations, SQR: Search queries, EXT: external sources, CHN: Channels browsed.

To additionally account for the use of each pathway, we quantify the contribution of each to ideological congeniality by factoring its frequency of use. We categorize the continuous slant score of each video into left-leaning, moderate, and right-leaning and classify as congenial a left-leaning video watched by *Very-liberal* and *Liberal* participants and right-leaning video watched by *Conservative* and *Very-conservative* participants. Plotting the extent of congeniality across pathways shows that the SBR recommendations lead to the highest congenial consumption at 35.1% , followed by the EXT pathway, with 32.7% of political videos watched by liberals and conservatives through this pathway being congenial, and by HPR with 21.7% congenial consumption. In short, the algorithmic pathways contribute to ideological congeniality on YouTube statistically significantly more than the user-driven pathways (i.e., combined SBR, HPR, and SQR at 62% versus combined EXT, and CHN at 38% (); $Z = 23.06$, $p < 0.01$ for a two proportion Z-test).

4.4.1.2. *Problematic Consumption.* To examine whether algorithmic or user-driven pathways lead users to consume problematic content, we use a combined list of 4,150 problematic channels [13, 10] and match consumption of each video from these channels to each of the analyzed pathways. Figure 4.2 shows the percentage of problematic videos consumed through each pathway and by participants' ideology. In contrast to congenial consumption, we see that `EXT` pathway drives the bulk of problematic consumption, at 46.1% across the sample, followed by `SBR` at 23.1%. When combining the pathways into algorithmic (i.e., `HPR`, `SBR`, and `SQR`) versus user-driven (`CHN` and `EXT`), we show that the latter lead to significantly more problematic consumption (user-driven 54.9% versus algorithmic 45.1% ; $Z = 3.71$, $p < 0.01$). In short, user-driven pathways (channels browsed and external sources) are the primary drivers of consumption of alternative and extremist channels, consistent with the findings of Chen et al. [6]. As found in past work [11, 6], we also show that *Conservative*, *Very-conservative* participants, and—surprisingly also the *Moderate* ones— had the highest consumption of content from problematic channels (i.e., 7.7%, 5.9%, and 5.9% respectively), over twice as high as the *Very-liberal* and *Liberal* participants (i.e., 2.9% and 2.8% respectively). We also find that the *Very-conservative* participants primarily watched problematic content from their homepage (35.6%) and sidebar recommendations (28.9%) in contrast to *Conservative* participants who more frequently came in from external sources (41.8%), with fewer consumption from their recommendations (`HPR` 20.6% `SBR` 18.2%). In fact, the bulk of the problematic consumption through external sources was driven by the *Moderate* users, followed by the *Very-liberal* users, who consumed problematic channels primarily through external sources (61.8% and 30% , respectively).

4.4.2. Analysis of Recommendations. So far, we show that ideologically congenial consumption is mostly driven by YouTube's algorithms whereas the consumption of problematic content is primarily driven by the users themselves. This evidence, similarly to [6, 11], centers on the videos consumed by users, i.e., what videos the users watched through which pathway. We take the next step and focus on the recommendations that the user received but did not click on. The core question regards whether—from the broad pool of homepage and sidebar recommendations—users *chose* more or less congenial (RQ2a) and more or less problematic (RQ2b) videos. If the user choices are *more* ideologically congenial and more problematic than the broader pool of recommendations made available to the user, it follows that it is the user's self-selection that drives these consumptions.

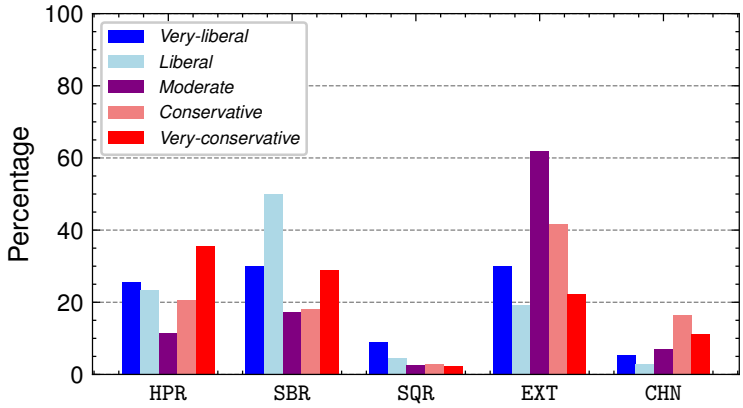


FIGURE 4.2. The proportion of problematic consumption by participants through each pathway. User pathways (EXT and CHN) lead to higher problematic consumption at 54.9% than algorithmic pathways (HPR, SBR, SQR) at 45.1%. The $\mathcal{N}$ values for *Very-liberal*, *Liberal*, *Moderate*, *Conservative*, and *Very-conservative* are 110, 68, 343, 165, 45, respectively.
HPR: Homepage recommendations, SBR: Sidebar recommendations, SQR: Search queries, EXT: External sources, CHN: Channels browsed.

	Homepage						Sidebar					
	Recommendations			Watched			Recommendations			Watched		
	Top-1	Top-8	All	Top-1	Top-8	All	Top-1	Top-8	All	Top-1	Top-8	All
Very-liberal	-0.29	-0.27	-0.21	-0.36	-0.34	-0.33	-0.31	-0.26	-0.26	-0.24	-0.27	-0.28
Liberal	-0.19	-0.16	-0.13	-0.29	-0.29	-0.25	-0.20	-0.18	-0.18	-0.11	-0.16	-0.16
Moderate	-0.05	0.01	-0.02	-0.13	-0.06	-0.06	-0.02	0.02	0.03	0.00	0.00	0.00
Conservative	0.44	0.41	0.23	0.37	0.40	0.42	0.29	0.30	0.31	0.22	0.31	0.31
Very-conservative	0.55	0.46	0.28	0.50	0.43	0.37	0.41	0.41	0.42	0.31	0.36	0.40

TABLE 4.3. Mean ideology estimates for homepage and sidebar recommendations and watched content, split by participant political ideology. Data is shown for the top-1 recommendation, top-8 recommendations, and all recommendations across both homepage and sidebar sections. Lower values indicate left-leaning content, while higher values indicate right-leaning content. The table highlights how the recommended content is less congenial than the actual content that participants end up watching but that there's a higher concentration of congenial content in the higher rankings.

4.4.2.1. *Congeniality in Recommendations.* Table 4.3 plots the mean slants of the videos recommended to the users on the homepage (column Homepage—Recommendations—All) and in the sidebar (column Sidebar—Recommendations—All). As seen, these recommendations are highly congenial with the users' ideologies. This is especially the case for the sidebar recommendations, where the mean slant is -0.26 (*median* = -0.42) and -0.18 (*median* = -0.39) for the *Very-liberal*

and *Liberal* participants, 0.31 (*median* = 0.23) and 0.42 (*median* = 0.46) for the *Conservative* and *Very-conservative* participants (column Sidebar—Recommendations—All in Table 4.3); with these being statistically significantly different from the videos recommended to *Moderate* participants.

However, as Table 4.3 also shows, the videos actually watched by the users (columns Homepage—Watched—All and Sidebar—Watched—All) from within those recommended are even *more* congenial than the recommendations (columns Homepage—Recommendations—All and Sidebar—Recommendations—All for homepage and sidebar, respectively). These differences are statistically significant for the *Very-liberal*, *Liberal*, and *Conservative* participants (see Table 4.4). This pattern is not as clear for the *Very-conservative* participants, where the mean slant of the videos watched is more congenial than the homepage recommendations and nearly identical as the sidebar recommendations.

	HPR	SBR
Very-liberal	0.15**	0.06**
Liberal	0.17**	0.05
Moderate	0.10**	0.05**
Conservative	0.17**	0.04
Very-conservative	0.11*	0.04

TABLE 4.4. KS-test to compare ideological estimates of videos watched with the videos recommended across ideologies. *HPR: Homepage recommendations, SBR: Sidebar recommendations, SQR: Search queries, EXT: External sources, CHN: Channels browsed.* Asterisks indicate statistical significance *: $p < 0.05$, ** $p < 0.01$.

These patterns clearly suggest a selection bias: despite the availability of ideologically diverse political recommendations on YouTube, participants self-select into watching content that is more congenial to their beliefs.

4.4.2.2. *Problematic Recommendations.* Does the same pattern of self-selection emerge for problematic content (RQ2b)? Table 4.5 shows the percentage of problematic homepage and sidebar recommendations alongside the percentage of problematic videos consumed by participants through these pathways. Here, the percentage is computed over the total number of political recommendations.

Table 4.5 presents the percentage of problematic recommendations received versus watched. As shown, a *higher* percentage of problematic videos was recommended to the participants on the homepage (column Homepage–Rec–All) than what was actually watched by the participants (column Homepage–Watched). For the *Very-conservative* participants, the percentages of problematic

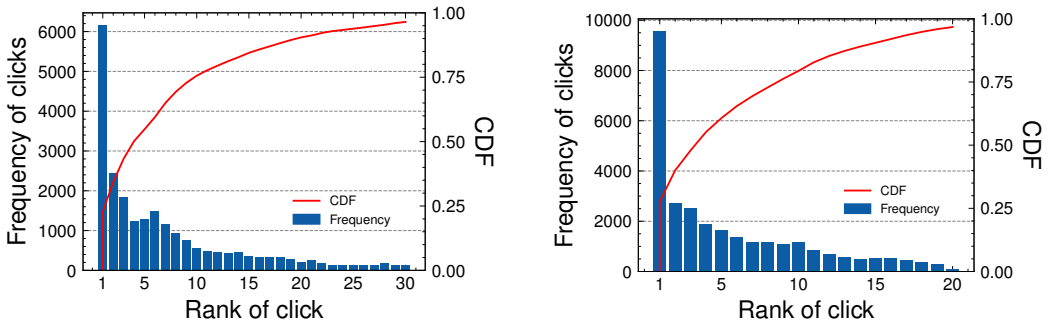
	Homepage				Sidebar			
	Recommendations			Watched	Recommendations			Watched
	Top-1	Top-8	All		Top-1	Top-8	All	
Very-liberal	3.3%	4.4%	2.5%	1.7%	2.8%	2.8%	2.8%	0.7%
Liberal	2.3%	2.7%	1.8%	1.5%	2.1%	2.7%	3.4%	0.5%
Moderate	2.3%	4.5%	3.0%	2.2%	3.3%	4.9%	5.0%	0.8%
Conservative	4.3%	9.1%	6.4%	3.4%	2.8%	5.3%	5.9%	1.2%
Very-conservative	5.0%	8.2%	6.1%	6.2%	4.1%	6.6%	6.6%	2.4%

TABLE 4.5. Percentage of problematic recommendations found in top-1 homepage and sidebar recommendations, top-8 homepage and sidebar recommendations, and all homepage and sidebar recommendations, and in the videos watched by the participants. Lower values in the watched show that participants watched fewer problematic recommendations than what they were recommended. Top-1 and top-8 columns are presented as fractions of number of videos in top-1 and top-8 whereas the other All and Watched columns are as a fraction of homepage, sidebar, and watched combined.

recommendations and watches are nearly the same. Similarly, the percentage of problematic sidebar recommendations (column Sidebar–Rec–All) was *higher* than what was actually selected and watched from these recommendations (column Sidebar–Watched). In short, participants do not watch as many problematic videos as are recommended to them by YouTube. Instead—as we show above—problematic content is mostly consumed through channels and off-platform sites. We do not observe differences in these patterns between ideologies, suggesting that all the participants consume fewer problematic videos than they are recommended.

4.4.3. Analysis of Ranking. Offering additional nuance, we account for the ranking of the recommended videos and its role in what users chose to watch. Does the ranking of recommendations, contribute to more or less (a) ideologically congenial and (b) problematic consumption (RQ3)? Recommended videos that are ranked higher in the users' feed are likely to be watched more often than videos which are ranked lower [106]. After all, users are unlikely to scroll down if something piques their interest towards the top. Then, the self-selection to more congenial content, found above, may be due to the algorithmic prioritization of certain contents.

Figure 4.3a indeed shows that 22.8% of homepage selections happen on the top-ranked video, 69.3% on the top-8, and 79.5% of the selections happen in the top-12 recommendations. As seen in Figure 4.3b, this is also true for sidebar recommendations, with 72.8% of clicks concentrated



(a) Homepage (b) Sidebar

FIGURE 4.3. Histogram (blue bars) showing raw counts and CDF (red line) showing the distribution of where users clicked in the homepage and sidebar recommendation rankings. We see that user clicks are concentrated in higher ranked content for both homepage and sidebar.

towards the top-8 higher ranked sidebar recommendations. We now examine whether there is a higher concentration of congenial and problematic videos in higher ranked recommendations than in others to account for the influence of rankings on consumption.

4.4.3.1. *Congeniality in Rankings.* Two clear patterns emerge when comparing the mean slant for top-1 and top-8 homepage and sidebar recommendations in Table 4.3. First, the videos the algorithm ranks higher for the users are significantly more congenial than all the recommendations on the homepage and, albeit less so, for the sidebar. Second, comparing the mean slant of the videos *consumed* in Table 4.3 (columns Homepage–Watched–All and Sidebar–Watched–All) with the mean slant of the top videos recommended (columns Homepage–Rec–Top-N and Sidebar–Rec–Top-N), the top homepage recommendations are significantly more congenial with the users' actual consumption than lower ranked recommendations. This holds particularly for the *Very-conservative* users, whose mean consumption is 0.37 (*median* = 0.32) but the top-1 and top-8 videos recommended on the homepage are 0.55 (*median* = 0.72) and 0.46 (*median* = 0.49) respectively. These differences are less pronounced for sidebar recommendations, where the slant of the top video prioritized for the users is similar to that of the top-8 and all sidebar videos recommended. In short, the user self-selection to congenial content may not necessarily reflect the users' agency itself but the fact that the algorithms prioritize very congenial content to the users, especially on the homepage.

4.4.3.2. *Problematic Rankings.* Table 4.5 shows that the percentage of problematic recommendations in the top-1 and especially top-8 homepage recommendations is slightly higher than in

all recommendations combined. These differences are more pronounced for the *Conservative* and *Very-conservative* participants, for whom problematic content is significantly more concentrated in the top-8 homepage and sidebar recommendations than across all homepage and sidebar recommendations. Despite this, the consumption of problematic content through these recommendations is infrequent, as seen earlier.

That is, although the problematic recommendations were, ever slightly so, concentrated towards the top of the homepage, participants did not consume those videos as much as they were recommended and overall consumption of problematic content was more frequent through user-driven pathways.

4.5. Discussion

This project aimed to determine whether it is the user or the algorithm that is primarily responsible for the consumption of both ideologically congenial and problematic content on YouTube. We combine granular data from frequent YouTube users on all the recommendations they received and the content they watched on the platform, all of which were collected using a web extension. We categorize consumption as either algorithmic or user-driven by identifying five pathways that led a user to a particular video and also the ranking of the videos in the homepage and sidebar recommendations. We use those data to pinpoint the role of algorithms and users in the consumption of both ideologically congenial content (based on a validated approach to video slant estimation) as well as content from extremist, conspiratorial, misinformative, and otherwise problematic YouTube channels compiled by past work [10, 88].

Our results provide a more precise understanding of the relationship between algorithms and users in shaping exposure to political and problematic content, and we offer two key findings.

First, our findings confirm that algorithmic ranking contributes meaningfully to ideological congeniality. When examining the pathways to congenial consumption, the algorithmic pathways (i.e., homepage and sidebar recommendations, search) lead to more congenial consumption than the user-driven pathways (i.e., channels browsed, external sources). That is, a higher percentage of ideologically congenial content is consumed when it is algorithmically recommended as opposed to when users browse certain channels or follow links from external sources. These echo the findings

of Haroon et al. [61] where they showed that even automated sock puppets following a strictly partisan viewing pattern also encountered ideologically congenial recommendations. Additionally, this disproportionality produces systematic differences in exposure across partisan groups. In this sense, the algorithm is not neutral: it nudges users toward content aligned with their prior preferences, constituting partisan information diets and reinforcing ideological divides [38, 98]. This evidence supports claims that recommender systems, even without malicious design, tend to heighten homogeneity because of engagement-based optimization. That said, we find that the users select the most ideologically aligned content from these recommendations. In other words, given a broader pool of recommended videos, the users actively *chose* videos that are yet more aligned with their political leanings and disregard those that are more diverse or cross-cutting. Although this indicates a self-selection bias, we also show that the recommendations that are prioritized and pushed to the top of the feed for the users are especially congenial in the case of political content. Because users tend to choose the top recommendations (i.e., 69.3% and 72.8% of clicks happen in the top-8 homepage and sidebar recommendations respectively) and that this is a well-documented behavior by prior work [106, 107, 108], this algorithmic prioritization greatly contributes to the patterns of congeniality observed. In short, this finding underscores the key role of recommendations and platform design that prioritizes recommendations that clearly align with the users' political biases. We note, however, that although the pattern of users selecting the higher-ranked recommendations holds throughout all the participants and it is unlikely users will scroll to click congenial content if it were ranked lower, further experimental designs are still needed to determine if the position of the recommendation indeed plays the dominant role in the user's decision to watch an (ideologically congenial) YouTube video.

Second, we find that problematic consumption is predominantly routed through user-driven pathways such as direct channel visits, and external referrals. These pathways reflect deliberate navigation choices, not algorithmic supply. This suggests that claims of autonomous "rabbit holes" induced by algorithms overstate the causal role of recommendation trails in explaining problematic viewing. Instead, problematic exposure seems to reflect demand-side selection, consistent with evidence that individuals often seek out ideologically extreme or conspiratorial sources on their own [20, 21].

This duality helps resolve a central debate in the literature. Critics emphasize the dangers of algorithmic radicalization, while defenders emphasize agency and demand. Our results suggest both perspectives capture important truths: algorithms do amplify congenial content, but problematic exposure reflects user-initiated actions. In other words, YouTube's recommendation system helps *constitute* partisan information diets, but it does not single-handedly drive widespread consumption of conspiratorial or extremist content. This nuanced attribution clarifies why prior work has reached seemingly contradictory conclusions [10, 11, 12, 6].

These findings carry clear implications for interventions. Because ranking systematically elevates congenial items, small adjustments to recommender objectives—such as incorporating diversity, quality, or civic value—could mitigate polarization without dramatically reducing engagement [21, 36]. At the same time, because problematic exposure flows mainly through user-driven pathways, interventions must also address demand-side dynamics. Policies that promote media literacy, reduce incentives for problematic creators, or curb cross-platform referrals may be more effective than focusing solely on algorithmic transparency or reform [35, 37].

Equally important, our results underscore the limits of one-size-fits-all claims about recommender systems. For ideologically congenial content, algorithmic supply matters greatly; for problematic content, user choice predominates. Future interventions must therefore be tailored to the type of harm being addressed: reducing echo chambers may require algorithmic tweaks, while combating conspiracy exposure may require tackling demand-side behaviors and community norms.

Having disentangled algorithmic and user contributions, the next step is to examine how interventions can modify these dynamics. Chapter 5 evaluates top-down approaches that alter algorithmic supply—such as changing ranking criteria or adjusting default slates—and bottom-up approaches that empower users to steer recommendations. Together, these strategies address both sides of the equation revealed here: algorithms that shape congenial exposure and users who drive problematic demand.

CHAPTER 5

Interventions on Social Platforms

5.1. Introduction

The preceding chapters demonstrated that YouTube's recommendation system amplifies *ideologically congenial content* and that user agency is central to exposure to *problematic channels*. These findings suggest that addressing harmful or undesirable outcomes requires interventions that target both the algorithmic supply of content and the ways users engage with it. In this chapter, we look at different intervention strategies that modify either the recommender system itself or the user experience to promote healthier information diets.

Prior research is limited to top-down interventions, which are designed from the perspective of the recommendation system designers. More specifically, prior work attempts to modify the recommendation algorithm (i.e., in-processing; e.g., architecture, optimization function) [109, 110, 111, 112, 113], the inputs to the algorithm (i.e., pre-processing; e.g., training dataset) [114, 115, 116, 117], or the output of the algorithm (i.e., post-processing; e.g., recommendation confidence) [118, 119, 120, 121] to debias recommendation systems. However, such mitigations can only be employed by the social media platforms themselves and they are wary of their bottom-line as these changes can impact their revenue streams. They are therefore not incentivized by applying these changes at a global level [122, 123, 124, 125, 126, 127, 128].

In contrast, bottom-up interventions that empower users directly are crucial for enhancing autonomy and agency as they can be deployed without relying on platform cooperation. The end goal of such interventions is to allow users control over their personalized recommendations and to steer the recommender system towards healthier information diets.

The rest of the chapter proceeds as follows. We first demonstrate a rudimentary de-biasing intervention to reduce the ideological congeniality of YouTube recommendations we observed in our earlier audit. We then operationalize our findings from this rudimentary intervention and our earlier

audits, and develop an intervention framework to promote healthier information diets on YouTube. We evaluate this framework through a large-scale user study and compare its effectiveness against user-centric nudges promoting better news consumption.

5.2. Background Interventions to Manipulate Recommendations

As our earlier audit demonstrated, there is a pressing need to mitigate ideological congeniality in YouTube recommendations to promote healthier news diets. Unfortunately, YouTube is not self-incentivized to fix its recommendation system, which is designed to primarily optimize user engagement [129]. In this section, we propose and systematically test a *bottom-up* intervention framework by designing a de-biasing approach, named CenterTube, that can be directly deployed to users. There are two main challenges in designing such an intervention in a principled manner. First, since YouTube's recommendation algorithm is proprietary[1] we do not possess exact knowledge of how the recommendation algorithm functions. Second, as a direct consequence, our intervention can only rely on *blackbox access* to YouTube's recommendation algorithm by querying it with an input (e.g., watching some videos) and observing the output (e.g., homepage recommendations). Given these challenges, we next describe the CenterTube framework to mitigate biased YouTube recommendations in a principled manner.

5.2.1. Proposed Approach.

5.2.1.1. *Overview.* We design a Reinforcement Learning (RL) [131] based intervention approach to mitigate ideologically congenial recommendations on YouTube. The intervention proceeds as follows. We train an RL model to minimize the recommendation of ideologically congenial YouTube content by promoting more ideologically diverse, dissimilar, and balanced content on the user's homepage as such diversity in information input is advantageous for social cohesion and citizen's attitudes, understanding, and knowledge [132]. The trained model *obfuscates* the user's watch history by injecting suitable intervention videos if a high degree of congeniality is detected on the user's homepage. We outline our RL methodology and key implementation details below.

[1]Google/YouTube have sporadically published high-level descriptions of their recommendation algorithms [130, 38, 129] but not in sufficient enough detail to be of any practical use.

5.2.1.2. *Injecting intervention videos.* Our intervention injects a fixed set of m videos at a time when the user is not actively using YouTube. We propose to train a model that operates on the user's YouTube homepage videos at any point in time and then chooses *intervention* videos to inject to debias the homepage if the homepage is ideologically biased (i.e., if its mean ideological slant lies in the *very-right* or *very-left* ideology bins). To define this analytically, let U denote the set of videos watched by the user before the intervention. Then, we can denote YouTube's recommendation algorithm as $\mathcal{A}$, which takes this watch history U as input, and outputs a homepage H consisting of a fixed number of k videos. Put simply, $H = \mathcal{A}(U)$.

At a high-level, the goal of the intervention is to *modify* U (the input to $\mathcal{A}$) by injecting m intervention videos, such that the obtained homepage H consists of more ideologically neutral, diverse, and/or dissimilar content so that it is balanced (i.e., mean ideology is close to 0).

5.2.1.3. *The goal of the RL model.* In the training phase, the RL model learns which videos to inject through *trial and error*. By training over time, it learns to minimize errors in decision-making, by maximizing a function known as *reward*. For our setting, we define this *reward* function to measure how ideologically balanced homepage recommendations are (i.e., how close the average ideology slant of these recommendations is close to 0).

To do this, the reward function takes as input a user's homepage, and outputs a score between -1 and 0, where 0 denotes a perfectly balanced/diverse homepage, and -1 denotes a *fully left* or *fully right* homepage. The goal of our intervention is to obtain a homepage where the recommended videos achieve a reward value of close to 0, indicating that there is a balance of *very-left* and *very-right* videos and/or that the majority of recommended videos are ideologically moderate. We can now define such a reward function analytically using the slant scores of videos (Equation 3.1).

Let the homepage be denoted as H. We define the reward function as $\mathcal{O}(H) = -|\overline{S}(H)|$ where $\overline{S}(H)$ denotes the mean slant scores of the set of homepage videos H. Again, the intervention injects a set of (*intervention* videos) so as to maximize $\mathcal{O}(H)$, whose maximum value of 0 represents a homepage comprising of videos with *mean* slant at center. If there are a large number of only left- or right-leaning videos, the reward value will be closer to -1. Mathematically, our intervention introduces a set of intervention videos I such that $\mathcal{O}(H') \to 0$, where $H' = \mathcal{A}(U \,\hat{\cup}\, I)$.[2] While the

[2]Note that the operator $\hat{\cup}$ here denotes an ordered union between the set of videos, as the order in which videos are watched is important.

reward function can be defined in different ways, our choice of $\mathcal{O}$ aims to minimize the mean ideology of a user's homepage. We select this reward function because it promotes ideological diversity. However, note that our approach is flexible enough to accommodate a variety of different such goals.

5.2.1.4. *Training the RL model.* Because we only have *blackbox access* to YouTube's recommendation algorithm $\mathcal{A}$, we can only see the output homepage recommendations *after* we inject a new video. Therefore, we design the learning framework as an *iterative* process where each time step represents the addition of a new intervention video to eventually make up the complete intervention video set I. Such a framework is also known as a Markov Decision Process (MDP) [133].

An MDP consists of the following components: *states*, *actions*, and the *reward function*, which we detailed above. For our setting, each *state* at a time step corresponds to the user's YouTube homepage at that time. Moreover, in an MDP, we can transition from one state to another using *actions*, and the goal is to choose actions such that we move to states where the reward is maximized. In our case, we choose from 3 actions: to either watch a 1) *very-left*-leaning video, 2) a *center*-leaning video, or a 3) *very-right*-leaning video.

In short, the training process for our setting is summarized as:

- For a learning *episode*, the model starts with the user's pre-intervention homepage as the first state.
- Depending on this state, the model picks a particular video (*very-left*, *center*, *very-right*) as the action to take.[3]
- The RL model goes back to the homepage and observes the recommendations to calculate the reward value using $\mathcal{O}$.
- Based on this reward obtained and the previous state, the model makes a decision for selecting the next action (i.e., the intervention video), and the process repeats from the first step.
- The learning *episode* concludes when m such videos have been introduced. We can hence train the RL model for many such episodes till it has learned to achieve desired reward, and can debias YouTube homepages effectively.

[3]The main motivation behind defining this action set is that each action should be able to send an explicit signal to the algorithm, to influence the recommendations. Thus, we only consider watching videos from the *very-left*, *very-right*, or *center* ideology bins, as opposed to *left* and *right* videos.

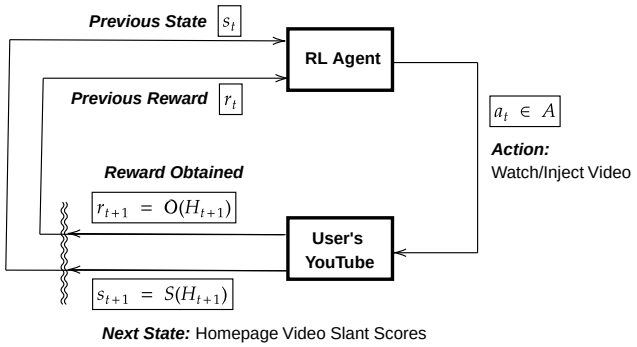


FIGURE 5.1. The proposed RL based approach.

- In this manner, this trained RL model has learnt an *optimal policy* that allows it to *map* each state (homepage) to the best possible action (video to inject) for that time step.

We will now describe some of these terms and formulations analytically.

States. At each time step t, the state s_t is defined as a k length vector of slant scores computed on the set of homepage H_t videos. We denote this as $s_t = S(H_t)$. Thus, our states encode the political ideology of the homepage at that time step.

Actions. At each time step t, the model interacts with the blackbox environment (the YouTube recommendation algorithm $\mathcal{A}$) by selecting one of the actions from the set of allowed actions $A = \{a_L, a_C, a_R\}$ where a_L, a_C, a_R refer to watching a *very-left*, *center*, or *very-right* video, respectively. As mentioned before, we opt for this action set since each of these actions will send an explicit signal to the recommendation algorithm $\mathcal{A}$.

Reward. At time step t, upon selecting a possible video injection a_t from A, the agent receives the set of next state observations s_{t+1} from the YouTube environment, as well as a scalar reward r_{t+1}. As described before, s_{t+1} is denoted as the slant scores of the homepage videos H_{t+1}. Since we want to learn a policy that maximizes the reward value, the immediate reward r_{t+1} is simply defined as $r_{t+1} = \mathcal{O}(H_{t+1}) = -|\overline{S}(H_{t+1})|$.

Model architecture. Using this formulated MDP, we learn an optimal policy π that maps each state to an action using RL algorithms. Specifically, we utilize the well-known Deep Q-Network [134] (DQN) RL algorithm. There are many benefits to using DQN and its extensions. First, the DQN algorithm is *simplistic*, and complements our MDP representing a simple/limited state-action

space. Second, the approach is *model-free* and *off-policy*, which does not require us to explicitly know the blackbox YouTube environment, and instead allows us to learn π using samples obtained from exploring the environment itself. Finally, combined with extensions such as Prioritized Experience Replay (PER) [135] and Dueling-DQN [136], the DQN architecture is stable and converges quickly. We employ these extensions in our experiments as well.[4] This description of the MDP and the RL based intervention framework is also illustrated in Figure 5.1.

We implement the intervention framework by training RL models for two types of users: left-leaning and right-leaning comprising of the *very-left* and *very-right* ideology bins respectively. We consider these two bins as they exhibit the most ideological congeniality in recommendations (as evidenced via the audit). Thus, given the homepage prior to the intervention, if the mean slant of the homepage (videos) lies in the *very-right* ideology bin we invoke the RL model for right-leaning users otherwise if it lies in the *very-left* ideology bin we use the RL model for left-leaning users.

There are many benefits to training *specialized* models as opposed to training a *general* model that can select intervention videos for users of any ideological leaning. A general model would have to first *learn* to distinguish between left-leaning and right-leaning users, whereas this information is encoded explicitly prior to training the specialized models. Note that this is trivial to do, since we can check whether $\overline{S}(H)$ is positive or negative prior to invoking the intervention. Other benefits to having specialized models include: (1) we can analyze each model's decisions more clearly since we know what they should intrinsically optimize for, (2) the MDP is less complex as each model is exposed to fewer states, and (3) models converge more quickly as the learning problem is simplified. This may be the biggest advantage – RL models are sample inefficient [137] and often require hundreds of thousands to millions of learning episodes to train well. As we cannot "simulate" YouTube offline, we have to resort to first training sock puppets to mimic real users, then watch videos and observe the homepage, which makes training an RL agent for such a large number of episodes not tractable. Thus, simplifying the MDP (state-action space) for specific types of users allows us to effectively train an RL model in fewer episodes.

[4]Because our state space consists of k length slant vectors, we use a Multi-layer Perceptron (MLP) Neural Network (2 layers with 64 neurons each) as our function approximator, as opposed to a more complex Convolutional Neural Network (CNN) used in the original paper [134].

Experiments for training and testing. We train the RL models by utilizing sock puppets as in Chapter 3. We begin a training episode by first training a sock puppet of the ideology corresponding to the particular RL model (i.e., left-leaning or right-leaning). As part of this process, each sock puppet "watches" 100 videos of the corresponding ideology; that is $|U| = 100$. Further, to analyze how many intervention videos are needed to debias the homepage efficaciously, we vary the parameter m from $25, 50, 75$, to 100 videos[5]. The training episode concludes when m videos have been injected. In this manner, the RL agent explores this state-action-reward space by consistently learning over 2500 episodes, and learns an optimal policy π that maps states to actions. We find that the model starts achieving desirable reward values after training for this duration.

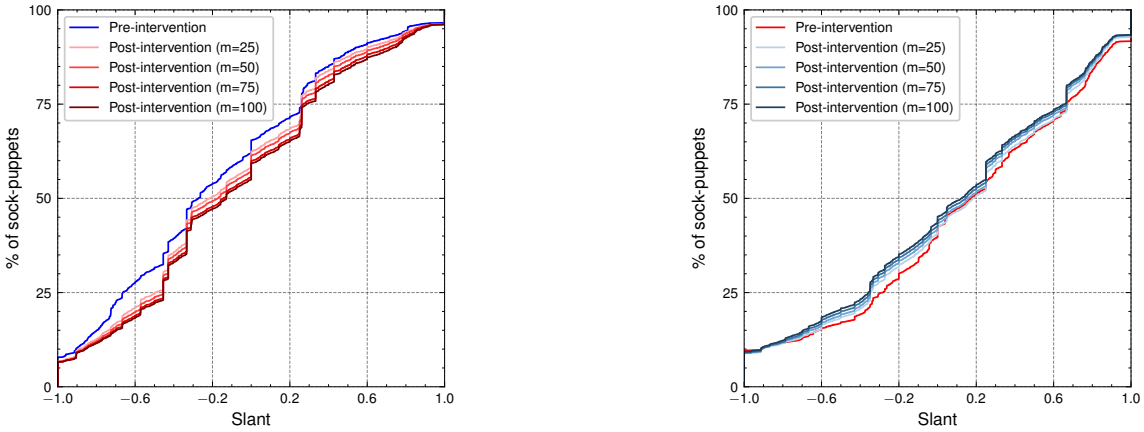
While we only train the models to debias homepage recommendations, we *test* the intervention out on both the homepage as well as up-next recommendations. In total, we employ 1894 sock puppets for these experiments which correspond to $1,589,105$ YouTube videos.

5.2.2. Results.

5.2.2.1. *Debiasing homepage recommendations.* We analyze whether the intervention framework debiases a user's homepages and how many intervention videos need to be injected. To this end, we collect homepage video recommendations for both *very-right* and *very-left* users (sock puppets), *before* and *after* the intervention and analyze the distributions of video slant scores seen by the sock puppets at both times to test whether they are *statistically significantly* different. Further, we also vary the value of m, i.e., the number of intervention videos injected from $25, 50, 75$ to 100 videos to ascertain how adding more intervention videos affects the extent to which the homepages is debiased.

The results are presented as Cumulative Distribution Functions (CDFs) for both the left-leaning and right-leaning model in Figure 5.2a and Figure 5.2b, respectively. The visual difference between the pre-intervention and post-intervention curves for both models indicates that the intervention does lead to a change in the ideological distribution of homepage videos. This *visual gap* between the pre-intervention and post-intervention curves is more pronounced for *very-left* sock puppets than for *very-right* sock puppets. Finally, increasing m leads to more debiasing as the pre-intervention and post-intervention curves become visually more distinct upon doing so.

[5]Note that we do not need to train separate RL models for different values of m as we can train a model for any m' and still utilize it for m intervention videos. Upon deployment, we can keep injecting videos sequentially till m videos are introduced.



(a) CDFs for left-leaning model.

(b) CDFs for right-leaning model.

FIGURE 5.2. CDFs showcasing the slant distribution for sock puppets pre-intervention and post-intervention (for different values of m, i.e., the # of intervention videos injected).

Next, we analyze the ideological leaning of recommended homepage videos before and after the intervention to assess if the intervention led to more ideologically diverse and balanced recommendations. Table 5.1 shows that for the left-leaning model, this is the case as the % of *very-left* videos decreases from 51.23% to 37.34% ($m = 100$), and both the *very-right* video % and *center* video % increase significantly– from 16.87% to 26.67% ($m = 100$), and from 31.90% to 35.99% ($m = 100$), respectively. For the right-leaning model, although there is a similar trend of increase in diverse content, the *very-right* video % does not decrease much (44.73% $\rightarrow$ 42.40%), and the *center* video % also decreases only slightly (31.66% $\rightarrow$ 28.41%) to account for the increase in *very-left* video % from 23.61% to 29.19% ($m = 100$). Thus, overall, more ideologically diverse and balanced content is recommended post the intervention for both models, even though the changes are less pronounced for right-leaning users.

5.2.2.2. *Debiasing up-next recommendations.* As mentioned before, while we train the RL models to debias homepage recommendations, we also analyze the effect of the intervention on up-next recommendations. To do so, we utilize the exposure metric (Equation 3.2). We use the same video seed as the ideological leaning of the sock puppet (*very-left* or *very-right*) and then compute exposure for recommendation trails of depth 20, pre- and post-intervention. As we found that increasing m

Model Type	Distribution	*very-left* video %	*very-right* video %	*center* video %
Left-leaning model	Pre-intervention	51.23	16.87	31.90
	Post-intervention ($m = 25$)	42.72	21.23	36.05
	Post-intervention ($m = 50$)	40.62	22.99	36.39
	Post-intervention ($m = 75$)	38.59	25.20	36.21
	Post-intervention ($m = 100$)	37.34	26.67	35.99
Right-leaning model	Pre-intervention	23.61	44.73	31.66
	Post-intervention ($m = 25$)	25.96	45.91	28.13
	Post-intervention ($m = 50$)	26.99	44.51	28.50
	Post-intervention ($m = 75$)	28.16	43.35	28.49
	Post-intervention ($m = 100$)	29.19	42.40	28.41

TABLE 5.1. Percentage of *very-left*, *very-right*, and *center* homepage videos encountered pre-intervention and post-intervention for both the left-leaning and right-leaning models.

improves debiasing of homepage recommendations, we utilize $m = 100$ for these experiments but note that other (smaller) values of m can also be used.

5.2.3. Takeaway. The intervention framework effectively debiases homepage recommendations for both left-leaning and right-leaning users, with more pronounced effects for left-leaning users. Increasing the number of intervention videos injected (m) leads to more balanced and ideologically diverse recommendations. While the use of the RL model is somewhat complex, the core finding that background interventions can effectively debias recommendations is significant. This approach can be adapted and operationalized for various types of mitigations beyond ideological congeniality. In the following chapter, we discuss the application of this intervention to promote healthier news diets on YouTube by deploying it to real users.

5.3. Nudging Interventions to Promote Healthy Information Diets

5.3.1. Motivation. The majority of Americans are active users of social media platforms. As those users navigate the online ecosystem, their attention and information consumption are largely driven by recommendation algorithms, which are designed to maximize engagement by recommending content that aligns with users' inferred interests and biases [1, 2, 3, 4]. These personalized recommendations drive between 75-95% of consumption on platforms [72, 6]. In this context, observers [8, 87] and scholars [11, 6, 13, 10] worry that algorithmic systems recommend like-minded political content and direct users to radical, conspiratorial, or otherwise problematic

information. Evidence to support these worries about what we refer to as *ideological bias* in algorithms is mixed at best [6, 11].

This project advances past work by tackling *interest bias* in recommendation algorithms, insufficiently recognized, under-studied, and crucially important for content exposure. Recommendation algorithms are designed to optimize the time the users spend on the platform by learning to recommend the content with which users are likely to engage [2, 3]. This feature of all recommendation systems is not on its own problematic. However, because most users go to platforms for entertainment, *not* for verified news and quality public affairs [27, 28, 29, 138], and because many engage with low-quality, sensationalist, and click-baity content [42, 139], algorithmic recommendations align with these interests and preferences [1, 140]. This, combined with the fact that algorithms actively direct users to entertainment and away from news [141] leads to closed loops of low-quality, non-news content consumption [142, 4, 28].

Under-consumption of public affairs has crucial democratic consequences. Low-information voters are vulnerable to irrelevant cues in the political environment [30], vote against their personal and group interests [31, 32, 33], and are more susceptible to populist, manipulative, and misinformative rhetoric [34]. In turn, news exposure leads to more informed citizens [143, 144, 145, 146, 147], increases opinion stability and voting in accordance with one's interests [148, 33], decreases beliefs in misinformation [36, 149, 150], enhances efficacy, tolerance, and the acceptance of democratic norms [151, 152], and leads to more equitable voting outcomes [153].

In addition, recommender systems may direct the more politically interested users to congenial, one-sided, and/or hyper-partisan political information, i.e., ideological bias in recommendation algorithms. Although few users inhabit echo chambers [138, 154, 155] or are put in extreme rabbit holes [6, 11], this small subset is consequential. Those users are more strongly partisan, more affectively polarized, more vocal and active in the political arena, and have a disproportionate influence on the democratic process [156, 157, 158]. Therefore, minimizing interest bias in recommendation algorithms and incentivizing greater consumption of verified news among citizens is of grave importance.

5.3.2. Proposed Interventions. One of the central challenges in mitigating online harms is that interventions must balance effectiveness with user autonomy. While top-down changes to recommender design can alter the supply of content, such approaches may be costly, opaque, or

politically contentious. By contrast, *nudging interventions* seek to modify the presentation of choices in ways that preserve freedom of selection while guiding users toward healthier or more diverse outcomes. Adapting the injection framework of CenterTube, we test whether providing *balanced* news input to the algorithm minimizes ideological congeniality in recommendations and consequently leads users to consume more diverse and cross-cutting news and political content. Because we establish the baseline slant of each user's news diet, from very liberal to very conservative, we can pinpoint daily over-time shifts in the slant of recommendations and exposures resulting from our intervention. In short, our intervention is built to increase news consumption while also minimizing ideological biases in information diets. Our expectation is that nudging the algorithm by altering a user's browsing history increases recommendations to and consumption of news and diverse news and political content.

Furthermore, we nudge individual users by reminding them of the personal and democratic benefits of news exposure. Many citizens avoid news, either actively or unintentionally [159]. They may be uninterested in politics, perceive it as irrelevant to their daily lives, or not make the connection between public affairs, governmental policies, and news events and their own and societal interest [160]. It follows that making politics more personally and collectively relevant, as well as making it easy for people to access news, could increase its consumption. We propose that nudging citizens to consume news by highlighting the benefits that news brings to the public and to oneself should encourage them to consume more news and political content.

This two-pronged approach—algorithmic and user nudge—is uniquely suited to disentangle two interrelated factors that underlie low news consumption: recommender systems and/or individual disinterest. If low news use is due to interest bias in algorithms (e.g., a sports fan is only recommended sports videos and is effectively secluded from news and public affairs information on platforms, [141]), then nudging the algorithm to recommend verified news could increase news consumption simply because more inventory is easily accessible. If, however, people avoid news mostly because they see it as irrelevant [159], reminding users that news benefits them and society could encourage greater news consumption.

In this project, we focus on YouTube, as it is one of the largest platforms, with 1.7 billion unique monthly visitors and 14.3 billion visits per month, more than Facebook, Wikipedia, Amazon, and

Instagram. YouTube is also the most popular platform in the U.S, used by 81% of the population, and has a steadily growing user base [25].

We conduct a systematic experiment using a web extension, which not only unobtrusively records all recommendations and information consumption on YouTube but also builds in two interventions: an algorithmic nudge and a user nudge. We deploy the extension in a longitudinal experiment with a sample of 2,142 frequent YouTube users recruited by YouGov, who installed the extension for one month. For the first week, the extension records all the recommendations (both homepage and up-next) and behaviors on the platform. In the next two weeks, the extension activates our interventions. In the *algorithmic nudge* treatment, the extension obfuscates user's YouTube watch history by unobtrusively playing videos from verified and balanced news sources, based on expert metrics. In the *user nudge* treatment, the extension changes the YouTube interface to integrate a banner reminding the users of the benefits of news exposure and a link to the YouTube News section. There are no changes to participants' YouTube experience in the control.

We examine the effects of our interventions—both during the treatment and over time after the interventions were deactivated—on (a) homepage and up-next recommendations and (b) the actual consumption of (i) news and (ii) political videos from outside of news channels (determined using our BERT-based classifier used earlier). We also test shifts in diversity of recommendations and consumption using our validated approach for estimating political slant of individual news and political videos (see [40]) *and also* to extreme, conspiratorial, and otherwise problematic channels from previously compiled lists [10, 88]. Lastly, we also rely on a post-test survey to test if increased news consumption has effects on self-reported outcomes (e.g., knowledge, participation, polarization, support for democratic norms. etc.).

We offer four key findings. First, the algorithmic nudge significantly and sustainably increases both recommendations and actual consumption of news on YouTube. In fact, we identify a reinforcing feedback loop between the algorithm and the users, in which recommendations drive news consumption more strongly than news consumption drives subsequent recommendations. Second, the algorithmic nudge also minimizes the ideological congeniality of recommendations and consumption, an effect that is most pronounced among conservatives and that persists after the intervention ends. Third, nudging individual users by changing the YouTube interface to integrate a banner

reminding the users of the benefits of news use has no observable effects on the consumption of (diverse) news and political content. Fourth, although the algorithmic nudge boosts the likelihood that users encounter and consume more (diverse and cross-cutting) news and political content, this increased news consumption has no effects on the tested attitudes and beliefs, null effects that do not vary with users' baseline news exposure or partisanship strength. Crucially, the intervention does not decrease the time spent and user engagement with YouTube, which makes it implementable by platforms concerned with engagement.

5.3.3. Experimental Design. We deployed a web extension among actual human users (who have dynamic watch histories and numerous on- and off-platform behaviors that sock puppets lack) in a month-long experiment on a sample of frequent YouTube users recruited by YouGov. Figure 5.5 outlines the design. The user experiment was conducted from November 2022 to January 2023. YouGov recruited American adults who visited YouTube more than once a week. 2,142 eligible participants were directed to the Chrome Store, where they were prompted to install our *ResearchTube* Chrome extension for four weeks. Participants were then randomly assigned to one of three experimental treatments: algorithmic nudge (35%), user nudge (35%), or control (30%).[6]

For the first week, the extension unobtrusively recorded all recommendations (homepage and up-next), exposures, and behaviors on the platform to establish a baseline. In weeks two and three, the extension additionally activates our interventions. In the *algorithmic nudge* treatment, the extension obfuscates user's YouTube watch history by unobtrusively playing/injecting news videos in the background every ten minutes when the user's browser is open. To ascertain ideological balance, for every four videos, one comes from left-leaning channels, two from centrist channels, and one from right-leaning channels based on expert metrics from Ad Fontes Media. In the *user nudge* treatment, the extension changes the YouTube interface to integrate a banner on the homepage and under a video a user is watching with a piloted prompt reminding the users of the benefits of news consumption. The extension randomly displayed a prompt from a list of four piloted prompts each time a user opened the YouTube homepage or watched a video.

[6]The assignment probabilities were kept uneven because higher dropout rates were anticipated amongst participants in the treatment groups as they would experience greater changes to their platform experience, such that the user nudge group was exposed to banners with the nudges and the algorithm nudge group experienced the greatest changes in their on-platform recommendations. Assigning more users to the treatment groups was to assure that we have sufficient statistical power to detect effects even in the light of potential higher dropout rates among these groups.

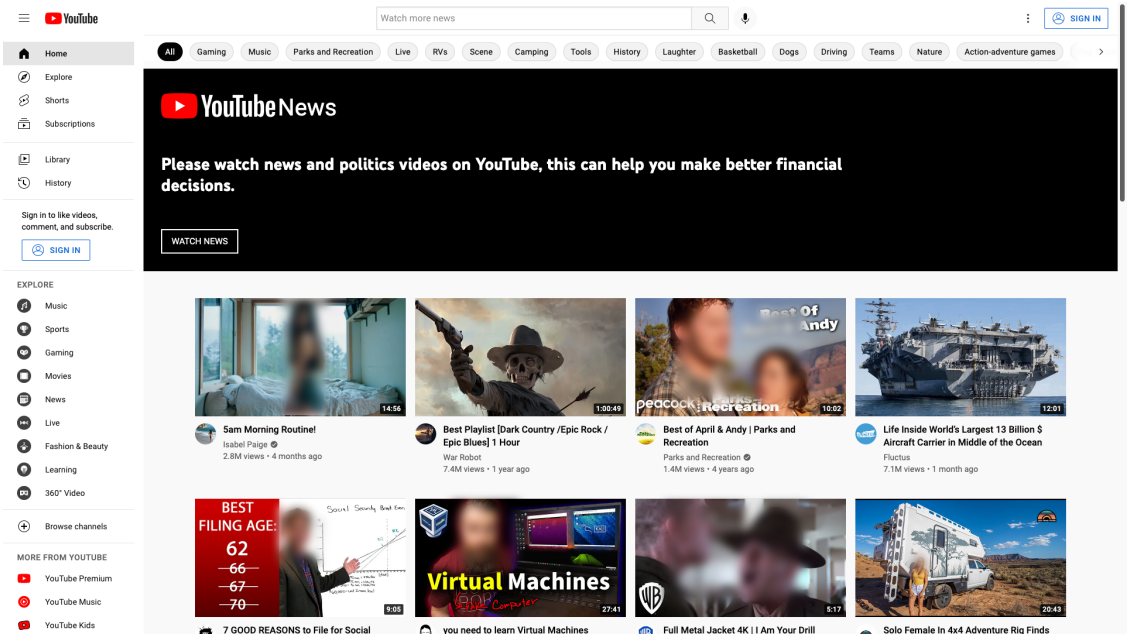


FIGURE 5.3. Example of YouTube Homepage in the User Nudge Treatment Condition

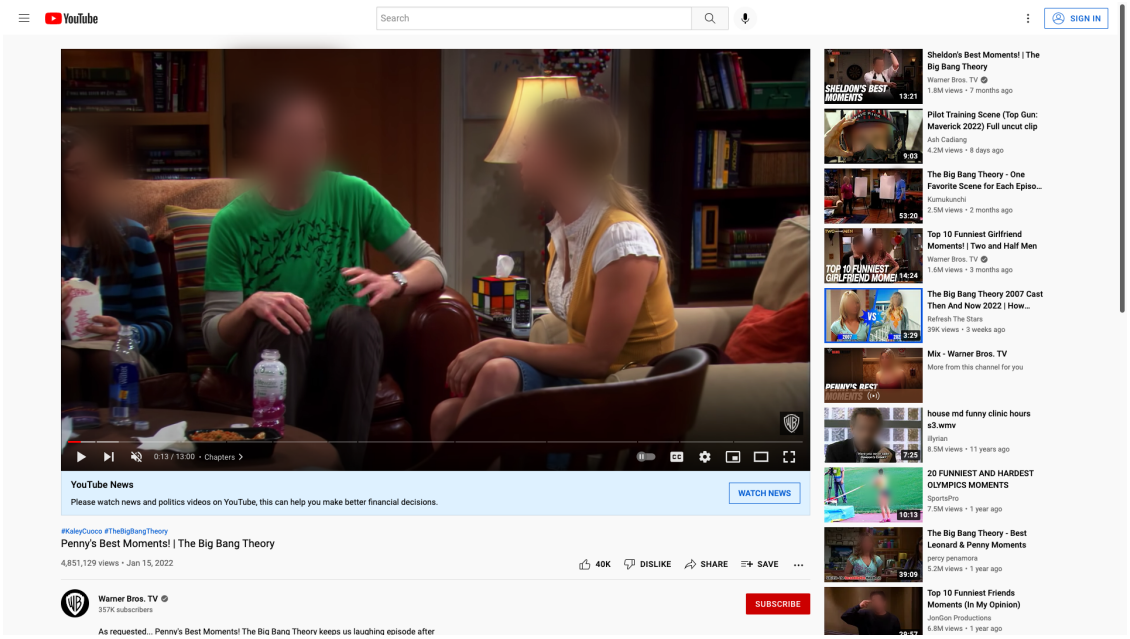


FIGURE 5.4. Example of Individual Video Page in the User Nudge Treatment Condition

Figures 5.3 and 5.4 show the appearance of the banner that appears for participants in the user nudge condition on the homepage and on any video page respectively. Following are the four prompts that were randomly selected from.

- *"In order to plan for your future, please watch news and politics videos on YouTube."*
- *"Please watch news and politics videos on YouTube, this can help you make better financial decisions."*
- *"Please watch news and politics videos on Youtube: A healthy democracy requires citizens who stay up to date with news and politics."*
- *"You can help strengthen our democracy, watch news and politics videos on YouTube today."*

Moreover, a button on the banner would take participants to YouTube News section and the YouTube search box suggested "Watch more news" (as can be seen in Figures 5.3 and 5.4). There are no modifications made to participants' YouTube experience in the control.

At the end of week three, participants who watched at least seven YouTube videos in weeks 1-3 and at least one video in weeks 2-3 (1,188) completed the post-test measuring key attitudes, cognitions, and behaviors (e.g., knowledge, participation, polarization, misinformation endorsement, support for democratic norms). These criteria ensured that participants were engaging with YouTube and were treated at least once. Out of 1,188 participants, 891 watched at least ten videos in weeks 2-3 and the median of videos watched during weeks 2-3 is 24. The interventions were terminated at the end of week 3. To capture any lasting effects, the extension continued to collect recommendations and users' on-platform behaviors for an additional full week. After four weeks, the extension was automatically uninstalled.

5.3.4. Data and Measures. The final sample includes participants who watched at least seven YouTube videos in weeks 1-3 and at least one video in weeks 2-3. This resulted in a total of 1,188 participants, with 430 in the algorithmic nudge treatment condition, 422 in the user nudge treatment condition, and 336 in the control condition. These inclusion criteria ensured that participants were engaging with YouTube and were treated at least once during the treatment weeks. The mean age of the final sample was 47.7 (SD = 16.5), with 46% males, 70% White, and 66% with a 4-year degree or some college. The majority (60%) identified as Democrats, 25% as Republicans, and 15% as Independents or other.

Overall, we collected data on 113,079 watched and 5,896,318 recommended videos, as well as the information on 15,218 YouTube channels participants browsed and the 14,956 search queries they used. Importantly, in Table 5.2 we show that neither the algorithmic nudge nor the user nudge

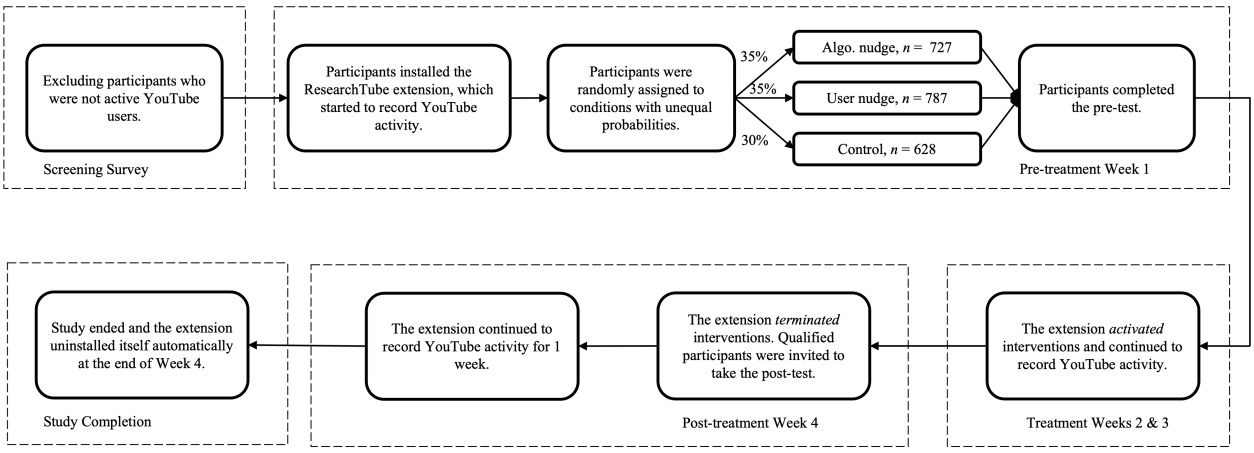


FIGURE 5.5. Overview of the research design. Participants were initially screened for their YouTube browsing habits and political ideology, and then suitable participants were asked to install the ResearchTube browser extension. The extension assigned either treatment or control and then monitored and administered interventions for the next four weeks.

treatments decrease YouTube usage (number of videos watched, number of days on YouTube, and time spent on the platform). In fact, users in the algorithmic nudge condition actually had 0.57 *more* active days on YouTube during the treatment weeks ($se = .12$, $p < .001$) than the control. This basic finding underscores the applicability of our intervention. In sum, the subtle sustained nature of the intervention and our ability to track YouTube recommendations and on-platform behaviors offer a unique opportunity to test the effects of algorithmic and user nudges over-time.

5.3.5. Results. First, we describe the percentage of news and political videos recommended and actually watched by users throughout the four weeks. As plotted in Figure 5.6, in week one, the percentage of *news* content watched (average: alg nudge 4.50%, user nudge 4.23%, control 4.22%) and recommendations (average: agl nudge 7.90%, user nudge 8.94%, control 8.26%) was relatively low across all participants. We observe a marked increase in recommendations to (16.42% on average) and viewership of (18.67% on average) news in the algorithmic nudge condition during the treatment weeks. News recommendations and viewership remained under 10% in the user nudge (7.67%, 4.26%) and control (7.69%, 3.59%). In week 4, when the intervention stopped, recommendations continued

	Videos watched per active day (W2-3)	Number of active days per week (W2-3)	Time spent per active day (W2-3)
Intercept	3.55 (1.17) **	1.16 (0.23) ***	4.92 (0.30) ***
Videos watched per active day (W1)	0.44 (0.02) ***	-	-
Number of active days (W1)	-	0.25 (0.02) ***	-
Average time spent per active day (W1)	-	-	0.40 (0.03) ***
Condition (algorithm)	0.78 (0.63)	0.57 (0.12) ***	-0.05 (0.09)
Condition (user)	0.77 (0.63)	0.08 (0.12)	0.17 (0.09)
N	1101	1188	1101
R^2	0.23	0.12	0.17

TABLE 5.2. OLS Regressions Predicting the Effects of Treatments on YouTube Engagement. *** < .001 ** < .01 * < .05. Time spent per active day (minutes) on YouTube (week 1 and weeks 2-3) were $\log_{x+1}$ transformed.

at 19.02% in the algorithmic nudge condition. Although news viewership declined to an average of 8.71%, it was still about double from the baseline in week 1.[7]

With regard to *political* content from outside news channels, Figure 5.6 shows that the percentage of recommended and watched political videos was consistently between 5% and 15% for all participants in week 1 (average recommendations: alg nudge 11.06%, user nudge 10.86%, control 9.56%; average consumption: alg nudge 11.54%, user nudge 9.74%, control 11.59%). There was no noticeable increase in recommendations to or watches of political videos in any of the conditions. These descriptives suggest that the algorithmic nudge—not the user nudge—increased both recommendations to and the actual watching of news but not political content outside of news channels.

We estimated the effects of the interventions, controlling for age, gender, education, race, partisanship, and the baseline number of news videos watched per active day in the pre-treatment week. The full models are plotted in Figure 5.7A. Confirming the descriptive plots above, the algorithmic nudge statistically significantly increased news consumption. Participants in the algorithmic nudge condition viewed 1.01 ($p < .001$) more news videos per active day during the treatment weeks compared to the control. This effect amounts to 2.59 times the average number of news videos all

[7]This drop may be due to the fact that for most participants, the last week of the experiment coincided with Christmas or New Years. Not only is it a slow news cycle but also people may be less inclined to watch news than other content. We tested if news videos were recommended higher up on users' homepages and in up-next recommendations during the treatment weeks than in week 4, finding no significant differences in the relative placement of news videos.

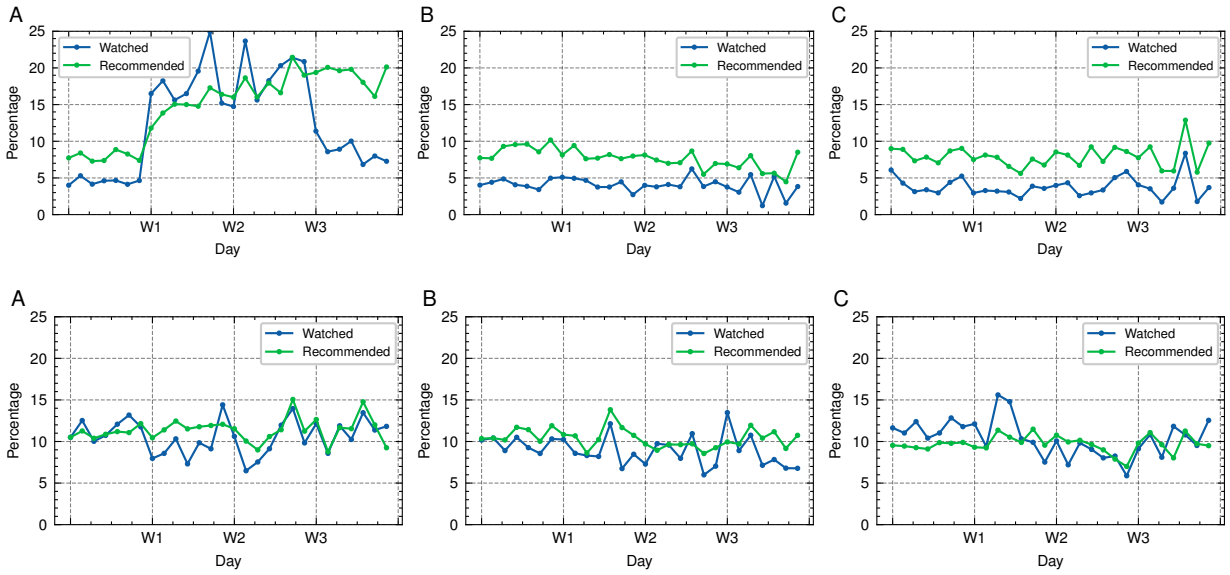


FIGURE 5.6. Average daily percentage of News (top row) and Political (bottom row) videos watched by and recommended to participants over total watches and recommendations. W1, W2, and W3 refer to the end of weeks 1, 2, and 3 respectively. The treatment ran between W1 and W3.
(A) Algorithmic nudge **(B)** User nudge **(C)** Control

participants watched per active day in week 1. If sustained for a year, users would consume 369 more news videos. This significant effect persisted even after the intervention terminated: in week 4, participants in the algorithmic nudge condition watched 0.25 ($p = .008$) more news videos per active day than the control. In turn, although there were slight increases in news watching in the user nudge condition, these effects were not statistically significant during the treatment weeks ($p = .258$) and post-treatment ($p = .236$). Also here, the nudges—whether the algorithmic or the user nudge—had no spillover effects on the consumption of political videos.

We ran parallel models predicting the *percentage* of news videos watched. Figure 5.7B additionally shows the effects. Participants in the algorithmic nudge condition watched 16% ($p < .001$) more news videos during the treatment weeks—nearly a fourfold increase compared to the average proportion of news videos all participants watched in week 1—and 7% ($p < .001$) more than the control in week 4. Again, there were no significant effects of the user nudge ($p = .516$ and .306 respectively). As before, the models find no effects of the two nudges on the consumption of political content. OLS

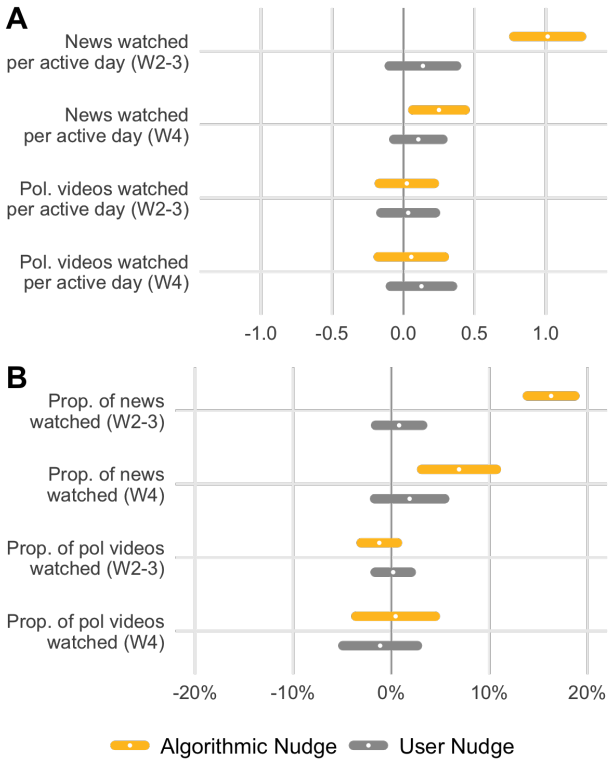


FIGURE 5.7. Effects of treatments on the consumption of news videos and political videos. Dots represent coefficients and horizontal bars represent 95% confidence intervals.

regression additionally showed these models with the algorithmic nudge treatment as the reference, finding that nudging the algorithm was significantly more effective than nudging the users.

Effects on Political Diversity of Recommendations and Consumption. Did the treatments make users' news diet more diverse and cross-cutting? We estimate the political slant of news and political videos in our dataset by analyzing the audience of each video on Twitter and the political landmarks these audiences follow. The approach, landmarks, score validation, and other details are presented in [40] and the code is available on Github. This approach estimated the slant for 85.2% of all news videos in our data and 72.2% of political videos.

Figure 5.8 shows the slant distribution of news and political videos recommended to (green plot) and actually consumed (blue plot) by (a) persons' ideology and (b) the phase of the intervention.[8] Four clear patterns emerge. First, each user group was recommended news and political videos

[8]Here, we combine news and political videos; the sparsity of those videos in the data, especially broken down by ideology and the phase of the intervention, would make separate estimates for news and political videos unstable.

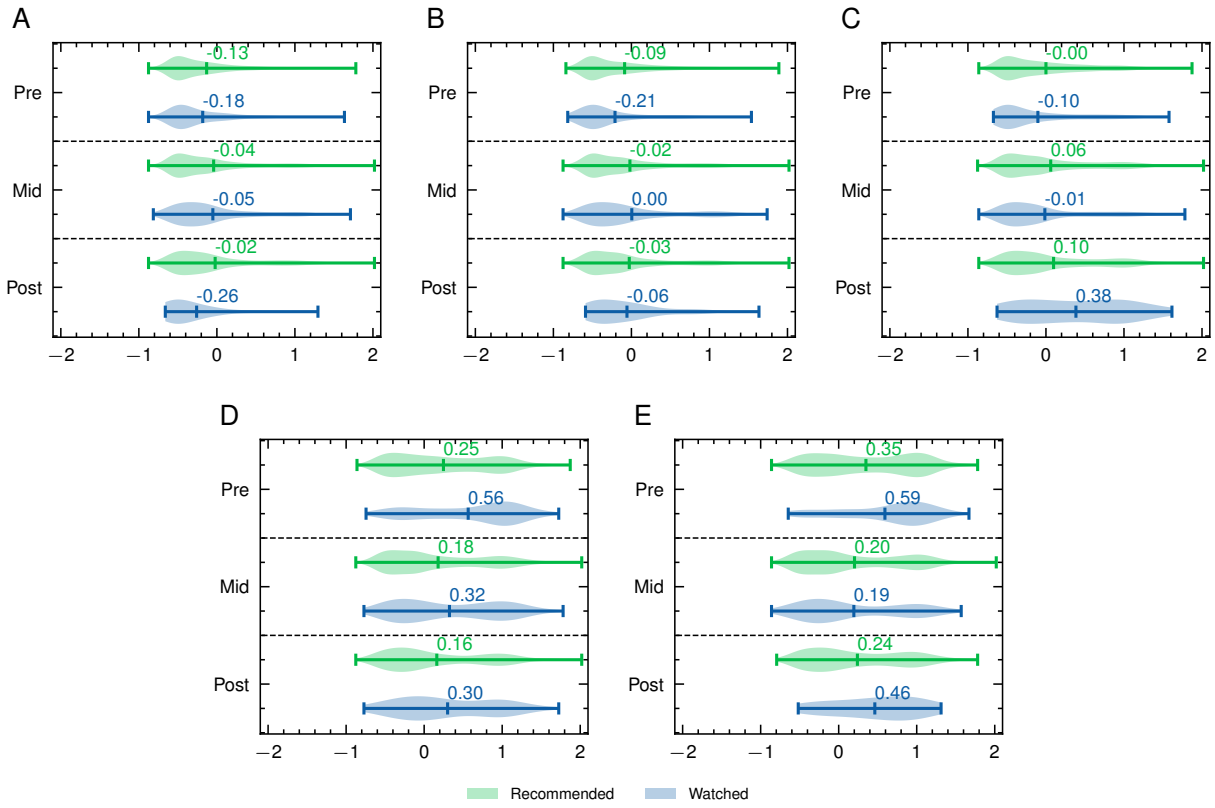


FIGURE 5.8. Political slant distribution of the videos recommended and watched, corresponding to the participants' self-reported political ideology. The green-colored distributions correspond to the slant of recommendations and blue-colored correspond to videos watched. The extreme ends represent the interquartile range and the values in-between are the mean slants of the distributions. We see that for the *very-conservative* and *conservative* participants, the recommendations weren't as partisan as the videos watched, suggesting that participants self-selected into their ideology rather than recommendations guiding them to do so.
(A) Very-liberal **(B)** Liberal **(C)** Moderate **(D)** Conservative **(E)** Very-conservative

consistent with users' ideological leaning. This congeniality was especially pronounced for *very-conservative* and *conservative* users (for whom the mean pre-intervention slant was +0.35 and +0.25, respectively). In contrast, the *very-liberal* and *liberal* participants' recommendations were also on the left yet closer to the neutral midpoint (-0.13 and -0.09, respectively)

Second, the algorithmic nudge noticeably and significantly moved the mean slant of the recommended news and political videos toward the center for all groups. This does not necessarily mean that the recommended news and political videos were strictly moderate, but that they were

diverse and cross-cutting, thus moving the mean slant toward 0. *Very-liberal* and *very-conservative* participants saw the highest shift, with an absolute difference of 0.11 in the mean slant pre- and post-intervention (*very-liberal* from -0.13 to -0.02; *very-conservative* from +0.35 to +0.24). Third, the intervention consequently made the participants' news and political diets *less* congenial. As seen in the middle panel of Figure 5.8, the mean slant of the videos watched was closer to 0 during intervention weeks than it was before. These differences between the pre- to mid-treatment distributions are statistically significant ($p < 0.05$) for both recommendations and consumption.

The last key finding in Figure 5.8 regards the videos recommended versus the videos watched. In almost all cases, the videos the users chose to consume were significantly more congenial than the videos they were recommended. This self-selection to congenial news and political videos was most pronounced for *conservative* and *very-conservative* participants. Pre-intervention, the average slant of the videos *recommended* to conservatives was 0.25 and 0.35 to very conservative participants, whereas the slant of the videos *watched* was 0.56 and 0.59, respectively. Post-intervention, these values were 0.16 and 0.24 (recommended) versus 0.30 and 0.46 (consumed). A similar—yet less pronounced—pattern of self-selection to congenial content also holds for the *very-liberal* and *liberal* users. Yet, caution is needed when interpreting the results from the post-intervention week, as attrition resulted in fewer active participants in each ideological group that week.

Effects of News Consumption on Attitudes and Beliefs. Lastly, we test if consumption of news influenced political participation, perceived accuracy of claims about current events, perceived polarization, among other outcomes assessed at the post-survey at the end of week 3. News exposure was operationalized as the total number of news videos consumed during treatment weeks *and* the difference between the number of news videos consumed per week during treatment weeks and in week one.[9] OLS regression models found that news exposure had consistent null effects on all survey outcomes and showed no statistically significant heterogeneous treatment effects on affective polarization and prioritizing partisan goals by partisanship strength, ideology, or ideology strength.

5.3.6. Discussion and Implications. Our nudging interventions demonstrate that subtle, interface-level changes can meaningfully alter what users see and choose without undermining overall

[9]The models included age, gender, education, race, partisanship, treatment assignment, pre-test values of the outcomes, and news exposure. Perceived accuracy of claims about current events was not measured in the pre-test.

engagement. The algorithmic nudge, which unobtrusively diversified participants' watch histories, substantially increased exposure to news and cross-cutting political content while maintaining usage levels (Table 5.2). By contrast, the user-facing prompts exerted weaker effects, underscoring the difficulty of altering entrenched behaviors through surface-level banners alone. These findings suggest that altering recommendation supply may be more effective than modifying demand through lightweight prompts [21, 36, 9]. Furthermore, nudging YouTube's algorithms by playing news videos in the background increases not only news recommendations but also the actual consumption of news. In fact, we identify a reciprocal loop, whereby recommendations drive exposure to news, which in turn promotes further recommendations, with the first link (recommendations $\rightarrow$ consumption) being stronger than the second link (consumption $\rightarrow$ recommendations). In other words, although users' on-platform behavior and content recommendations are intrinsically related, the algorithms seem to matter more to users' information consumption than what the users consume matters to the algorithms, at least in the tested context. This feedback loop amplifies the effects of the algorithmic nudge. Indeed, we observed that a week after the treatment, users in the algorithmic nudge condition were still being recommended and consuming more news than those in the control group. In short, a simple back-end tweak to the algorithms can promote greater engagement with public affairs.

We find no effects on recommendations and exposure to political videos from outside news channels (e.g., a vlogger discussing politics or a celebrity endorsing a candidate). This could be because YouTube algorithms pull users away from public affairs [141]. While our algorithmic treatment managed to counteract this algorithmic design (after all, we injected news into the users' watch history), it may not have been strong enough to generate a spillover effect on recommendations to political videos.

Providing politically balanced news input to the algorithm (recall that our algorithmic intervention was programmed to unobtrusively watch videos from across the political spectrum) leads to more diverse and cross-cutting recommendations and consequently minimizes congeniality in users' political and news exposure. During the treatment, recommendations in the algorithmic nudge condition shifted toward more moderate and cross-cutting, which led to more diverse news consumption. We emphasize, however, that the users were consuming significantly more congenial videos than they were being recommended. Although making news easily available does lead users to consume news,

the users exercise agency by self-selecting into specific news that aligns with their ideology. This reinforces a key insight: algorithmic changes can broaden exposure, but user agency continues to play a decisive role in shaping ultimate consumption patterns. In this sense, nudges operate most effectively as complements rather than substitutes for recommender-level reforms.

Nudging users by reminding them that news is beneficial to democracy and personally relevant fails to increase news consumption. This may be due to the limited participants' attention to the nudge, which did not change the behavior that it was theorized and predicted to induce. Only 12.7% participants in the user nudge condition clicked on the *"Watch News"* button and there was no marked increase in search queries for news. In addition, the null effects from the user nudge may be because the participants already reported believing that news consumption is important. In the pre-test survey, participants' average agreement with the idea that watching news and politics is crucial for society and themselves was high (M=4.05 out of 5) and there was no significant increase in this self-reported agreement in the user nudge condition compared to those in the algorithmic (p = .959) and control (p = .060) conditions.

Although the algorithmic intervention increased news consumption, this increase had no corresponding effects on a range of survey outcomes, including belief accuracy, political knowledge, or affective polarization neither in the aggregate nor across subgroups of participants. This finding adds to the growing evidence that while (algorithmic) interventions can alter users' on-platform exposure, this exposure does not influence attitudes or beliefs [21, 75]. We emphasize that this does not mean that promoting news consumption is futile or that the pronounced effects on recommendations and actual news exposure in our algorithmic condition are meaningless. Media effects accumulate over time and our two-week intervention may have been too short or insufficiently strong for these effects to emerge. It is also possible that increased news exposure may have affected outcomes that we did not measure, such as political efficacy [161], subsequent news seeking [162], or increased awareness of social problems [163]. Future studies should include these additional outcomes. We also encourage cross-platform and comparative work testing whether similar results would be detected on other platforms and in other countries.

Our findings carry important implications. Enhancing the accessibility of news on social media not only encourages its consumption but also fosters a reinforcing loop during user-algorithm interactions.

This indicates a substantial amount of power wielded by proprietary platform algorithms. Platforms could easily implement an intervention analogous to ours, e.g., up-ranking and prioritizing verified news content, yet some platforms do just the opposite, redirecting users from news [141] or down-ranking or outright banning such content [164]. Although platforms argue that users are not interested in news [165], evidence suggests that users want to see more accurate and educational information [166] and enjoy following news accounts on social media [150]. We additionally show that increasing recommendations to news does not decrease users' engagement with YouTube.

The broader democratic implications of nudging deserve emphasis. By increasing exposure to news and reducing the congeniality of political diets, even modestly, nudging interventions can help mitigate polarization risks while preserving freedom of choice. Importantly, these interventions avoid heavy-handed censorship or radical platform redesign, instead empowering users with subtle cues and re-weighting existing supply. As such, they represent a pragmatic strategy for platforms seeking to improve information ecosystems without sacrificing engagement metrics or provoking strong political backlash.

CHAPTER 6

Conclusion

This dissertation examined the role of recommender systems in shaping information consumption, with a particular focus on YouTube. Across a series of audits and interventions, we combined large-scale data collection with experimental methods to disentangle how algorithms and users interact to produce exposure to *ideologically congenial content* and *problematic channels*, and to test strategies for mitigating undesirable outcomes. Together, these studies advance empirical, methodological, and theoretical understanding of how recommender systems operate and how their effects may be meaningfully governed.

6.1. Summary of Findings

Chapter 3 introduced a scalable auditing framework for examining the behavior of recommender systems through trained synthetic sock puppets. By systematically exposing these sock puppets to curated viewing histories, the audit isolates algorithmic effects while approximating realistic user behavior. The results reveal consistent ideological congeniality in YouTube's recommendations: the system tends to surface like-minded content aligned with each persona's political profile, extending recent work on algorithmic reinforcement. Congeniality is most pronounced for very-right sock puppets, which receive disproportionately aligned recommendations. At the same time, evidence for algorithmic radicalization remains limited. Increases in ideological extremity across up-next trails are small, and while a non-trivial fraction of sock puppets encounter problematic channels, such exposure is neither systematic nor dominant.

Chapter 4 shifts from algorithmic supply to realized behavior. Using fine-grained telemetry from real users, the analysis decomposes political and problematic video consumption into algorithmic and user-driven pathways. A central finding is that congenial consumption flows primarily through algorithmic pathways (homepage, sidebar, and search). In contrast, problematic content is disproportionately accessed via user-driven mechanisms such as channel pages, and external referrals.

This distinction reconciles competing claims in prior work: while recommender systems do reinforce political leanings, users themselves account for much of the demand for conspiratorial or harmful content.

Chapter 5 examined whether ranking dynamics can be steered through bottom-up interventions. A month-long field experiment demonstrates that introducing small amounts of balanced, verified news content into users' watch histories increases recommendations to news, improves the ideological diversity of recommended content, and raises downstream news consumption. Importantly, this occurs without reducing overall engagement suggesting that modest perturbations to recommendations can yield meaningful outcomes. This finding highlights the changeability of algorithmic outputs and offers a practical pathway for healthier and positive content consumption.

Across the three empirical chapters, the dissertation shows that (1) YouTube's algorithm amplifies ideological congeniality but is not a dominant driver of extremist exposure; (2) harmful content consumption is primarily user-driven; and (3) targeted, minimal interventions can meaningfully improve information diets. These findings underscore that algorithmic and user-driven forces must be understood jointly rather than in isolation.

6.2. Implications

6.2.1. Theoretical Implications. The results nuance theoretical accounts of algorithmic influence. Claims of deterministic, self-reinforcing radicalization are not supported by the evidence: while YouTube's system reinforces political alignment, it does not strongly escalate ideological extremity. This suggests that theories of algorithmic agency should model amplification as bounded and asymmetric rather than runaway. Moreover, the empirical separation between politically congenial content and democratically problematic content highlights the need for theoretical frameworks that distinguish preference reinforcement from motivated harmful information-seeking. The dissertation also contributes to emerging discussions of "interest bias," where engagement-optimized systems systematically down-rank verified news in favor of entertainment or emotionally charged content. The intervention study demonstrates that small shifts in inferred interests can alter these outcomes, suggesting that interest-based distortions are neither fixed nor inevitable.

6.2.2. Methodological Implications. Methodologically, the dissertation introduces several tools that expand the capacity of researchers and regulators to study recommender systems. The trained sock-puppet framework provides a middle ground between naive synthetic accounts and costly user panels, enabling scalable yet interpretable audits. The video-level ideology estimator enables scalable, fine-grained classification beyond channel-level labeling. The pathway decomposition method offers a generalizable technique for attributing exposure to algorithmic versus user-driven mechanisms, an increasingly important task for accountability and governance. Finally, the field-experiment infrastructure demonstrates that researchers can conduct realistic interventions without platform cooperation, opening new avenues for audits.

6.2.3. Policy and Governance Implications. The findings have direct implications for discussions of platform accountability and regulatory design. First, they show that independent audits are feasible and capable of revealing systematic patterns that would otherwise remain opaque. Regulators may adopt similar frameworks to evaluate compliance with transparency mandates or algorithmic auditing requirements. Second, because problematic content consumption is largely user driven, governance initiatives must consider not only algorithmic ranking but also platform affordances, user incentives, and cross-platform diffusion mechanisms. Third, the effectiveness of bottom-up interventions suggests opportunities for designing user-facing tools, public-interest media partnerships, or civic-technology infrastructures that improve exposure to credible information without relying on platform cooperation. These results contribute to ongoing debates about the appropriate distribution of responsibility between platforms, policymakers, and the public.

6.3. Future Directions

Building on the contributions of this dissertation, several avenues for future research emerge:

- **Cross-platform audits.** Extend the auditing framework to TikTok, Instagram Reels, Facebook, and other emerging platforms to assess whether similar amplification, congeniality, and interest-bias patterns arise across distinct algorithmic architectures.
- **Longitudinal and real-time auditing.** Develop infrastructures to monitor recommender systems continuously, enabling detection of algorithmic drift, policy changes, and emergent risks that may not be visible in one-shot audits.

- **Hybrid causal attribution designs.** Combine sock-puppet audits, consumption pathways, surveys, and causal inference techniques to more precisely characterize bidirectional user and algorithmic feedback loops.
- **LLM-driven synthetic personas.** Explore the use of large language models to generate realistic, consistent user personas for audits, which may capture nuanced human-like engagement patterns while reducing reliance on sensitive user data.
- **Heterogeneous intervention effects.** Investigate how the effectiveness of bottom-up interventions varies across ideological groups, media diets, political sophistication, and demographic characteristics, and design targeted mechanisms accordingly.
- **Modeling interest bias and prioritization.** Examine how recommender systems infer user interests, prioritize content categories, and implicitly trade off engagement with informational quality, particularly in their treatment of verified news.
- **Governance and transparency tools.** Develop practical audit and monitoring tools that journalists, regulators, and civil society organizations can use to evaluate ranking systems and assess compliance with transparency or accountability mandates.
- **Cross-platform diffusion and pathways.** Integrate pathway analysis with cross-platform trace data to understand how political and problematic content spreads across interconnected ecosystems, including messaging apps and social networks.
- **Normative and ethical frameworks.** Advance theoretical work on the ethics of auditing and intervention, including issues of consent, autonomy, and the proper governance of systems capable of shaping public attention and civic discourse.
- **Designing for civic resilience.** Collaborate with news organizations, educators, and civic-technology actors to develop scalable interventions that increase exposure to credible information and strengthen democratic resilience in algorithmic environments.

6.4. Closing Remarks

This dissertation provides an integrated account of how recommender systems and user behavior jointly structure online information consumption. By combining algorithmic audits, disentangling content consumption into individual pathways, and field experiments, it advances understanding

of the mechanisms behind political and problematic exposure and demonstrates that information ecosystems can be improved through targeted, low-cost interventions. As digital platforms continue to mediate civic life, developing rigorous, transparent, and accountable methods for understanding and governing recommender systems will remain essential for supporting informed citizenship and safeguarding democratic processes.

Bibliography

[1] Michael A DeVito. From editors to algorithms: A values-based approach to understanding story selection in the Facebook news feed. *Digital journalism*, 5(6):753–773, 2017.

[2] T. A. Kendall, M. R. Cohler, M. E. Zuckerberg, Y. F. Juan, R. K. X. Jin, J. M. Rosenstein, and C. M. Palihapitlya. US Patent No. 12/193,705, 2009. US Patent No. 12/193,705.

[3] Timothy Kendall and Ding Zhou. Leveraging information in a social network for inferential targeting of advertisements, oct 7 2010. US Patent App. 12/419,958.

[4] Kjerstin Thorson, Kelley Cotter, Mel Medeiros, and Chankyung Pak. Algorithmic inference, political interest, and exposure to news and politics on facebook. *Information, Communication & Society*, 24(2):183–200, 2021.

[5] WSJ Staff. Inside Tiktok's Highly Secretive Algorithm. *https://www.wsj.com/video/series/inside-tiktoks-highly-secretive-algorithm/investigation-how-tiktok-algorithm-figures-out-your-deepest-desires/*, 2021.

[6] Annie Y Chen, Brendan Nyhan, Jason Reifler, Ronald E Robertson, and Christo Wilson. Subscriptions and external links help drive resentful users to alternative and extremist youtube channels. *Science Advances*, 9(35):eadd8080, 2023.

[7] Silvia Milano, Mariarosaria Taddeo, and Luciano Floridi. Recommender systems and their ethical challenges. *Springer AI & SOCIETY*, 35(4):957–967, 2020.

[8] Zeynep Tufekci. Youtube, the great radicalizer. *The New York Times*, 10(3):2018, 2018.

[9] Eli Pariser. *The Filter Bubble: What the Internet Is Hiding from You.* Penguin Press, 2011.

[10] Manoel Horta Ribeiro, Raphael Ottoni, Robert West, Virgílio AF Almeida, and Wagner Meira Jr. Auditing Radicalization Pathways on YouTube. In *ACM Conference on Fairness, Accountability, and Transparency (FAT*)*, 2020.

[11] Homa Hosseinmardi, Amir Ghasemian, Aaron Clauset, Markus Mobius, David M Rothschild, and Duncan J Watts. Examining the consumption of radical content on youtube. *Proceedings*

of the National Academy of Sciences, 118(32):e2101967118, 2021.

[12] Kostantinos Papadamou, Savvas Zannettou, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Michael Sirivianos. "It is just a flu": Assessing the Effect of Watch History on YouTube's Pseudoscientific Video Recommendations. In *AAAI Conference on Web and Social Media*, volume 16, pages 723–734, 2022.

[13] Mark Ledwich, Anna Zaitsev, and Anton Laukemper. Radical bubbles on YouTube? Revisiting algorithmic extremism with personalised recommendations. *First Monday*, 27(12), Dec. 2022.

[14] William Hart, Dolores Albarracín, Alice H Eagly, Inge Brechan, Matthew J Lindberg, and Lisa Merrill. Feeling Validated Versus Being Correct:A Meta-Analysis of Selective Exposure to Information. *Psychological bulletin*, 135(4):555–588, 2009.

[15] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, 27(1):415–444, 2001.

[16] Kasey Lobaugh, Jeff Simpson, and Lokesh Ohri. Navigating the new digital divide: Capitalizing on digital influence in retail. *Deloitte Digital*, 2015.

[17] Eslam Hussein, Prerna Juneja, and Tanushree Mitra. Measuring Misinformation in Video Search Platforms: An Audit Study on YouTube. In *ACM Conference On Computer-Supported Cooperative Work And Social Computing (CSCW)*, 2020.

[18] Leonardo Sanna, Romano Salvatore, Corona Giulia, and Agosti Claudio. YTTREX: crowd-sourced analysis of YouTube's recommender system during COVID-19 pandemic. In *International Conference on Information Management and Big Data*, 2020.

[19] Kostantinos Papadamou, Savvas Zannettou, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Michael Sirivianos. "How over is it?" Understanding the Incel Community on YouTube. In *ACM Conference On Computer-Supported Cooperative Work And Social Computing (CSCW)*, 2020.

[20] Megan A Brown, James Bisbee, Angela Lai, Richard Bonneau, Jonathan Nagler, and Joshua A Tucker. Echo chambers, rabbit holes, and algorithmic bias: How youtube recommends content to real users. *Available at SSRN 4114905*, 2022.

[21] Brendan Nyhan, Jaime Settle, Emily Thorson, Magdalena Wojcieszak, Pablo Barberá, Annie Y Chen, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Drew Dimmery, et al. Like-minded

sources on facebook are prevalent but not polarizing. *Nature*, 620(7972):137–144, 2023.

[22] Frank Newport and Andrew Dugan. Partisan Differences Growing on a Number of Issues. Gallup. https://news.gallup.com/opinion/polling-matters/215210/partisan-differences-growing-number-issues.aspx, 2017.

[23] Shanto Iyengar, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra, and Sean J. Westwood. The Origins and Consequences of Affective Polarization in the United States. *Annual Review of Political Science*, 22(1):129–146, 2019.

[24] Nathan P Kalmoe and Lilliana Mason. *Radical American partisanship: Mapping violent hostility, its causes, and the consequences for democracy*. University of Chicago Press, 2022.

[25] Brooke Auxier and Monica Anderson. Social Media Use in 2021. Pew Research. `https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/`, 2021.

[26] Ashley Rodriguez. YouTube's recommendations drive 70% of what we watch. Quartz. `https://qz.com/1178125/youtubes-recommendations-drive-70-of-what-we-watch`, 2018.

[27] C. McClain, R. Widjaya, G. Rivero, and A. Smith. The behaviors and attitudes of U.S. adults on Twitter, 2021.

[28] Chris Wells and Kjerstin Thorson. Combining big data and survey techniques to model effects of political content flows in Facebook. *Social Science Computer Review*, 35(1):33–52, 2017.

[29] Meta. Widely viewed content report: What people see on Facebook, 2022. `https://transparency.fb.com/data/widely-viewed-content-report/`.

[30] Larry M Bartels. Uninformed votes: Information effects in presidential elections. *American journal of political science*, pages 194–230, 1996.

[31] Richard R Lau, David J Andersen, and David P Redlawsk. An exploration of correct voting in recent US presidential elections. *American Journal of Political Science*, 52(2):395–411, 2008.

[32] Richard R Lau and David P Redlawsk. *How voters decide: Information processing in election campaigns*. Cambridge University Press, 2006.

[33] Scott L Althaus. Information effects in collective preferences. *American Political Science Review*, 92(3):545–558, 1998.

[34] Jason Brennan. Against democracy. In *Against Democracy*. Princeton University Press, 2017.

[35] Jack Bandy. Problematic machine behavior: A systematic literature review of algorithm audits. *Proceedings of the acm on human-computer interaction*, 5(CSCW1):1–34, 2021.

[36] Sacha Altay, Rasmus Kleis Nielsen, and Richard Fletcher. News can help! the impact of news media and digital platforms on awareness of and belief in misinformation. *The International Journal of Press/Politics*, 29(2):459–484, 2024.

[37] Hadi Askari, Anshuman Chhabra, Bernhard Clemm von Hohenberg, Michael Heseltine, and Magdalena Wojcieszak. Incentivizing news consumption on social media platforms using large language models and realistic bot accounts. *PNAS nexus*, 3(9):pgae368, 2024.

[38] Paul Covington, Jay Adams, and Emre Sargin. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*, pages 191–198, 2016.

[39] Karan Vombatkere, Sepehr Mousavi, Savvas Zannettou, Franziska Roesner, and Krishna P Gummadi. Tiktok and the art of personalization: investigating exploration and exploitation on social media feeds. In *Proceedings of the ACM Web Conference 2024*, pages 3789–3797, 2024.

[40] Magdalena Wojcieszak, Rong-Ching (Anna) Chang, and Ericka Menchen-Trevino. Political content and news are polarized but other content is not in youtube watch histories. *Journal of Quantitative Description: Digital Media*, 3, Nov. 2023.

[41] Cody Buntain, Richard Bonneau, Jonathan Nagler, and Joshua A Tucker. Youtube recommendations and effects on sharing across online social platforms. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–26, 2021.

[42] Steve Rathje, Jay J Van Bavel, and Sander Van Der Linden. Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences*, 118(26):e2024292118, 2021.

[43] D. Klug, Yiluo Qin, Morgan C. Evans, and Geoff F. Kaufman. Trick and please. a mixed-method study on user assumptions about the tiktok algorithm. In *13th ACM Web Science Conference 2021*, 2021.

[44] Juliane A. Lischka and Marcel Garz. Clickbait news and algorithmic curation: A game theory framework of the relation between journalism, users, and platforms. *New Media & Society*, 2021.

[45] David Chavalarias, Paul Bouchaud, and Maziyar Panahi. Can few lines of code change society? beyond fack-checking and moderation: how recommender systems toxifies social networking sites. *arXiv preprint arXiv:2303.15035*, 2023.

[46] Eytan Bakshy, Solomon Messing, and Lada A Adamic. Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132, 2015.

[47] Christopher A Bail, Lisa P Argyle, Taylor W Brown, John P Bumpus, Haohan Chen, MB Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37):9216–9221, 2018.

[48] Pablo Barberá, John T Jost, Jonathan Nagler, Joshua A Tucker, and Richard Bonneau. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological science*, 26(10):1531–1542, 2015.

[49] Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–36, 2017.

[50] Engin Bozdag and Jeroen Van Den Hoven. Breaking the filter bubble: democracy and design. *Ethics and information technology*, 17(4):249–265, 2015.

[51] Mark Alfano, Amir Ebrahimi Fard, J Adam Carter, Peter Clutton, and Colin Klein. Technologically scaffolded atypical cognition: The case of youtube's recommender system. *Synthese*, 199:835–858, 2021.

[52] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. Bias and Debias in Recommender System: A Survey and Future Directions. *ACM Trans. Inf. Syst.*, 41(3), feb 2023.

[53] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry*, 22(2014):4349–4357, 2014.

[54] Smitha Milli, Micah Carroll, Yike Wang, Sashrika Pandey, Sebastian Zhao, and Anca D Dragan. Engagement, user satisfaction, and the amplification of divisive content on social media. *PNAS nexus*, 4(3):pgaf062, 2025.

[55] George Beknazar-Yuzbashev, Rafael Jiménez-Durán, Jesse McCrosky, and Mateusz Stalinski. Toxic content and user engagement on social media: Evidence from a field experiment. Technical report, CESifo Working Paper, 2025.

[56] Naijia Liu, Xinlan Emily Hu, Yasemin Savas, Matthew A Baum, Adam J Berinsky, Allison JB Chaney, Christopher Lucas, Rei Mariman, Justin de Benedictis-Kessner, Andrew M Guess, et al. Short-term exposure to filter-bubble recommendation systems has limited polarization effects: Naturalistic experiments on youtube. *Proceedings of the National Academy of Sciences*, 122(8):e2318127122, 2025.

[57] Ferenc Huszár, Sofia Ira Ktena, Conor O'Brien, Luca Belli, Andrew Schlaikjer, and Moritz Hardt. Algorithmic amplification of politics on twitter. *Proceedings of the national academy of sciences*, 119(1):e2025334119, 2022.

[58] Sarmad Chandio, Muhammad Daniyal Pirwani Dar, and Rishab Nithyanand. How audit methods impact our understanding of youtube's recommendation systems. *Proceedings of the International AAAI Conference on Web and Social Media*, 18(1):241–253, May 2024.

[59] Annie Y. Chen, Brendan Nyhan, Jason Reifler, Ronald E. Robertson, and Christo Wilson. Exposure to Alternative & Extremist Content on YouTube. Anti Defamation League. `https://www.adl.org/resources/reports/exposure-to-alternative-extremist-content-on-youtube`, 2021.

[60] Martin Hilbert, Arti Thakur, Pablo M Flores, Xiaoya Zhang, Jee Young Bhan, Patrick Bernhard, and Feng Ji. 8–10% of algorithmic recommendations are 'bad', but... an exploratory risk-utility meta-analysis and its regulatory implications. *International Journal of Information Management*, 75:102743, 2024.

[61] Muhammad Haroon, Magdalena Wojcieszak, Anshuman Chhabra, Xin Liu, Prasant Mohapatra, and Zubair Shafiq. Auditing youtube's recommendation system for ideologically congenial, extreme, and problematic recommendations. *Proceedings of the National Academy of Sciences*, 120(50):e2213020120, 2023.

[62] Martin Hilbert, Drew P Cingel, Jingwen Zhang, Samantha L Vigil, Jane Shawcroft, Haoning Xue, Arti Thakur, and Zubair Shafiq. # bigtech@ minors: Social media algorithms personalize minors' content after a single session, but not for their protection. *Available at SSRN 4674573*,

2023.

[63] David A. Broniatowski, Joseph R. Simons, Jiayan Gu, Amelia M. Jamison, and Lorien C. Abroms. The efficacy of facebook's vaccine misinformation policies and architecture during the covid-19 pandemic. *Science Advances*, 9(37):eadh2132, 2023.

[64] Muhammad Ali, Piotr Sapiezynski, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. Discrimination through optimization: How facebook's ad delivery can lead to skewed outcomes. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–30, 2019.

[65] Wen Chen, Diogo Pacheco, Kai-Cheng Yang, and Filippo Menczer. Neutral bots probe political bias on social media. *Nature Communications*, 12(1):5580, Sep 2021.

[66] Maximilian Boeker and Aleksandra Urman. An empirical investigation of personalization factors on tiktok. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 2298–2309, New York, NY, USA, 2022. Association for Computing Machinery.

[67] Ronald E Robertson, Jon Green, Damian J Ruck, Katherine Ognyanova, Christo Wilson, and David Lazer. Users choose to engage with more partisan news than they are exposed to on google search. *Nature*, pages 1–7, 2023.

[68] Allison Chaney, Brandon Stewart, and Barbara Engelhardt. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *ACM Conference on Recommender Systems*, 2018.

[69] M Shamim Hossain and Ghulam Muhammad. Emotion recognition using deep learning approach from audio–visual emotional big data. *Information Fusion*, 49:69–78, 2019.

[70] Michal Kosinski, David Stillwell, and Thore Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the national academy of sciences*, 110(15):5802–5805, 2013.

[71] Yilun Wang and Michal Kosinski. Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of personality and social psychology*, 114(2):246, 2018.

[72] WSJ Staff. Inside Tiktok's Highly Secretive Algorithm. https://www.wsj.com/video/series/inside-tiktoks-highly-secretive-algorithm/investigation-how-tiktok-algorithm-figures-out-your-deepest-desires/, 2021.

[73] Juan Carlos Medina Serrano, Orestis Papakyriakopoulos, and Simon Hegelich. Dancing to the partisan beat: A first analysis of political communication on tiktok. In *Proceedings of the 12th ACM Conference on Web Science*, pages 257–266, 2020.

[74] Andrew Guess, Benjamin Lyons, Jacob M Montgomery, Brendan Nyhan, and Jason Reifler. Fake news, facebook ads, and misperceptions. *Democracy Fund*, 2019.

[75] Andrew M Guess, Neil Malhotra, Jennifer Pan, Pablo Barberá, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Drew Dimmery, Deen Freelon, Matthew Gentzkow, et al. Reshares on social media amplify political news but do not detectably affect beliefs or opinions. *Science*, 381(6656):404–408, 2023.

[76] Xudong Yu, Hennes-Michel Barnehl, and Magdalena Wojcieszak. Promoting news and political information consumption on social media with self-interest and public-interest nudges. *Information, Communication & Society*, pages 1–30, 2025.

[77] Rajvardhan Oak, Muhammad Haroon, Claire Wonjeong Jo, Magdalena Wojcieszak, and Anshuman Chhabra. Re-ranking using large language models for mitigating exposure to harmful content on social media platforms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 894–908, 2025.

[78] Rahul Bhargava, Anna Chung, Neil S Gaikwad, Alexis Hope, Dennis Jen, Jasmin Rubinovitz, Belén Saldías-Fuentes, and Ethan Zuckerman. Gobo: A system for exploring user control of invisible algorithms in social media. In *Companion Publication of the 2019 Conference on Computer Supported Cooperative Work and Social Computing*, pages 151–155, 2019.

[79] Geza Kovacs, Zhengxuan Wu, and Michael S Bernstein. Rotating online behavior change interventions increases effectiveness but also increases attrition. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–25, 2018.

[80] Smita Ghosh, Pramita Das, Sneha Ghosh, and Diptaraj Sen. Detection of clickbait content spreaders on online social networks. In *2022 5th International Conference on Information and Computer Technologies (ICICT)*, pages 23–28. IEEE, 2022.

[81] Zewdie Mossie and Jenq-Haur Wang. Vulnerable community identification using hate speech detection on social media. *Information Processing & Management*, 57(3):102087, 2020.

[82] Fabio Del Vigna12, Andrea Cimino23, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the first Italian conference on cybersecurity (ITASEC17)*, pages 86–95, 2017.

[83] Liang Wu, Fred Morstatter, Kathleen M Carley, and Huan Liu. Misinformation in social media: definition, manipulation, and detection. *ACM SIGKDD explorations newsletter*, 21(2):80–90, 2019.

[84] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36, 2017.

[85] George Beknazar-Yuzbashev, Rafael Jiménez Durán, Jesse McCrosky, and Mateusz Stalinski. Toxic content and user engagement on social media: Evidence from a field experiment. *Available at SSRN 4307346*, 2022.

[86] Roger McNamee. Platforms Must Pay for Their Role in the Insurrection. WIRED. `https://www.wired.com/story/opinion-platforms-must-pay-for-their-role-in-the-insurrection/`, 2021.

[87] Kevin Roose. The Making of a YouTube Radical. The New York Times. `https://www.nytimes.com/interactive/2019/06/08/technology/youtube-radical.html`, 2019.

[88] Mark Ledwich and Anna Zaitsev. Algorithmic extremism: Examining youtube's rabbit hole of radicalization. *First Monday*, 25(3), Feb. 2020.

[89] Pablo Barberá. Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data. *Political Analysis*, 23(1):76–91, 2015.

[90] Jon Green, Stefan McCabe, Sarah Shugars, John Harrington, Hanyu Chwe, Luke Horgan, Shuyang Cao, and David Lazer. Curation bubbles: Domain versus url level analysis of partisan news sharing on social media, Sep 2021.

[91] Huyen Le, Zubair Shafiq, and Padmini Srinivasan. Scalable News Slant Measurement Using Twitter. In *AAAI Conference on Web and Social Media*, volume 11, 2017.

[92] Vanessa Otero. Ad fontes media's first multi-analyst content analysis ratings project, 2019.

[93] Yadolah Dodge. *Spearman Rank Correlation Coefficient*, pages 502–505. Springer New York, New York, NY, 2008.

[94] Stefan McCabe and Jon Green. Updated twitter-based ideal points, 2021. GitHub repository.

[95] How Google uses cookies. Google. https://policies.google.com/technologies/cookies?hl=en-U, 2023.

[96] Manage your recommendations and search results. Google. https://support.google.com/youtube/answer/6342839, 2021.

[97] Cristos Goodrow. On YouTube's recommendation system. https://blog.youtube/inside-youtube/on-youtubes-recommendation-system/, 2021. Accessed: 2024-08-15.

[98] James Davidson, Benjamin Liebald, Junning Liu, Palash Nandy, Taylor Van Vleet, Ullas Gargi, Sujoy Gupta, Yu He, Mike Lambert, Blake Livingston, and Dasarathi Sampath. The YouTube Video Recommendation System. In *ACM Conference on Recommender Systems*, 2010.

[99] Galen Stocking, Patrick van Kessel, Michael Barthel, Katerina Eva Matsa, and Maya Khuzam. Many Americans Get News on YouTube, Where News Organizations and Independent Producers Thrive Side by Side. Pew Research, 2020.

[100] YouTube. Youtube community guidelines. https://www.youtube.com/howyoutubeworks/policies/community-guidelines/, 2021.

[101] Aaron Mackey and Kurt Opsahl. Van Buren is a Victory Against Overbroad Interpretations of the CFAA, and Protects Security Researchers. EFF. https://www.eff.org/deeplinks/2021/06/van-buren-victory-against-overbroad-interpretations-cfaa-protects-security, 2021.

[102] Casey Fiesler, Nathan Beard, and Brian C Keegan. No robots, spiders, or scrapers: Legal and ethical regulation of data collection methods in social media terms of service. In *AAAI conference on web and social media*, volume 14, pages 187–196, 2020.

[103] Sandvig v. Barr – Challenge to CFAA Prohibition on Uncovering Racial Discrimination Online. ACLU. https://www.aclu.org/cases/sandvig-v-barr-challenge-cfaa-prohibition-uncovering-racial-discrimination-online, 2019.

[104] Naomi Gilens and Jamie Williams. Federal Judge Rules It Is Not a Crime to Violate a Website's Terms of Service. EFF. https://www.eff.org/deeplinks/2020/04/federal-judge-rules-it-not-crime-violate-websites-terms-service, 2020.

[105] Stephan Lewandowsky, Ronald E. Robertson, and Renee DiResta. Challenges in understanding human-algorithm entanglement during online information consumption. *Perspectives on Psychological Science*, 19(5):758–766, 2024. PMID: 37427579.

[106] Andrew Collins, Dominika Tkaczyk, Akiko Aizawa, and Joeran Beel. Position bias in recommender systems for digital libraries. In Gobinda Chowdhury, Julie McLeod, Val Gillet, and Peter Willett, editors, *Transforming Digital Worlds*, pages 335–344, Cham, 2018. Springer International Publishing.

[107] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, Filip Radlinski, and Geri Gay. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Trans. Inf. Syst.*, 25(2):7–es, apr 2007.

[108] Mark Glick, Greg Richards, Margarita Sapozhnikov, and Paul Seabright. How does ranking affect user choice in online search? *Review of Industrial Organization*, 45:99–119, 2014.

[109] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, 2017.

[110] Shi-Yong Chen, Yang Yu, Qing Da, Jun Tan, Hai-Kuan Huang, and Hai-Hong Tang. Stabilizing Reinforcement Learning in Dynamic Environment with Application to Online Recommendation. In *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018.

[111] Xiangyu Zhao, Liang Zhang, Zhuoye Ding, Long Xia, Jiliang Tang, and Dawei Yin. Recommendations with negative feedback via pairwise deep reinforcement learning. In *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018.

[112] Robin Burke, Nasim Sonboli, and Aldo Ordonez-Gauger. Balanced Neighborhoods for Multi-sided Fairness in Recommendation. In *Conference on Fairness, Accountability and Transparency*, 2018.

[113] Golnoosh Farnadi, Pigi Kouki, Spencer K. Thompson, Sriram Srinivasan, and Lise Getoor. A fairness-aware hybrid recommender system, 2018.

[114] Jiangxing Yu, Hong Zhu, Chih-Yao Chang, Xinhua Feng, Bowen Yuan, Xiuqiang He, and Zhenhua Dong. Influence Function for Unbiased Recommendation. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020.

[115] Dugang Liu, Pengxiang Cheng, Zhenhua Dong, Xiuqiang He, Weike Pan, and Zhong Ming. A General Knowledge Distillation Framework for Counterfactual Recommendation via Uniform Data. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020.

[116] Bashir Rastegarpanah, Krishna P Gummadi, and Mark Crovella. Fighting Fire with Fire: Using Antidote Data to Improve Polarization and Fairness of Recommender Systems. In *ACM International Conference on Web Search and Data Mining*, 2019.

[117] Paula Gordaliza, Eustasio Del Barrio, Gamboa Fabrice, and Jean-Michel Loubes. Obtaining Fairness using Optimal Transport Theory. In *International Conference on Machine Learning*, 2019.

[118] Preetam Nandy, Cyrus DiCiccio, Divya Venugopalan, Heloise Logan, Kinjal Basu, and Noureddine El Karoui. Achieving fairness via post-processing in web-scale recommender systems. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 715–725, New York, NY, USA, 2022. Association for Computing Machinery.

[119] Faisal Kamiran, Asim Karim, and Xiangliang Zhang. Decision Theory for Discrimination-Aware Classification. In *IEEE International Conference on Data Mining*, 2012.

[120] Moritz Hardt, Eric Price, and Nati Srebro. Equality of Opportunity in Supervised Learning. *Advances in neural information processing systems*, 29:3315–3323, 2016.

[121] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On Fairness and Calibration. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, 2017.

[122] Harald Steck. Training and testing of recommender systems on data missing not at random. In *ACM SIGKDD International Conference on Knowledge Discovery and Data mining (KDD)*, 2010.

[123] Benjamin M Marlin and Richard S Zemel. Collaborative prediction and ranking with non-random missing data. In *ACM Conference on Recommender Systems*, 2009.

[124] Jiawei Chen, Can Wang, Martin Ester, Qihao Shi, Yan Feng, and Chun Chen. Social recommendation with missing not at random data. In *IEEE International Conference on Data*

Mining (ICDM), 2018.

[125] Yixin Wang, Dawen Liang, Laurent Charlin, and David M. Blei. The deconfounded recommender: A causal inference approach to recommendation, 2018.

[126] Yuta Saito. Asymmetric Tri-training for Debiasing Missing-Not-At-Random Explicit Feedback. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020.

[127] Wenlong Sun, Sami Khenissi, Olfa Nasraoui, and Patrick Shafto. Debiasing the human-recommender system feedback loop in collaborative filtering. In *World Wide Web Conference*, 2019.

[128] Francesco Fabbri, Francesco Bonchi, Ludovico Boratto, and Carlos Castillo. The Effect of Homophily on Disparate Visibility of Minorities in People Recommender Systems. In *AAAI International Conference on Web and Social Media (ICWSM)*, 2020.

[129] Minmin Chen, Alex Beutel, Paul Covington, Sagar Jain, Francois Belletti, and Ed H Chi. Top-k off-policy correction for a reinforce recommender system. In *ACM International Conference on Web Search and Data Mining*, 2019.

[130] Renjie Zhou, Samamon Khemmarat, and Lixin Gao. The impact of YouTube recommendation system on video views. In *ACM SIGCOMM conference on Internet measurement*, pages 404–410, 2010.

[131] Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*. MIT press, 2018.

[132] Natali Helberger, Kari Karppinen, and Lucia D'acunto. Exposure diversity as a design principle for recommender systems. *Information, Communication & Society*, 21(2):191–207, 2018.

[133] Richard Bellman. A Markovian Decision Process. *Journal of mathematics and mechanics*, 6(5):679–684, 1957.

[134] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning, 2013.

[135] Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay, 2015.

[136] Ziyu Wang, Tom Schaul, Matteo Hessel, Hado Hasselt, Marc Lanctot, and Nando Freitas. Dueling network architectures for deep reinforcement learning. In *International Conference on Machine Learning*, 2016.

[137] Denis Yarats, Amy Zhang, Ilya Kostrikov, Brandon Amos, Joelle Pineau, and Rob Fergus. Improving Sample Efficiency in Model-Free Reinforcement Learning from Images. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

[138] Magdalena Wojcieszak, Andreu Casas, Xudong Yu, Jonathan Nagler, and Joshua A Tucker. Most users do not follow political elites on Twitter; those who do show overwhelming preferences for ideological congruity. *Science Advances*, 8(39):eabn9418, 2022.

[139] Xudong Yu, Magdalena Wojcieszak, and Andreu Casas. Partisanship on social media: In-party love among american politicians, greater engagement with out-party hate among ordinary users. *Political Behavior*, 46(2):799–824, 2024.

[140] Wilbert Samuel Rossi, Jan Willem Polderman, and Paolo Frasca. The closed loop between opinion formation and personalized recommendations. *IEEE Transactions on Control of Network Systems*, 9(3):1092–1103, 2021.

[141] Sandy Huang and Tong Yang. Auditing entertainment traps on YouTube: How do recommendation algorithms pull users away from news. *Political Communication*, pages 1–19, 2024.

[142] Kjerstin Thorson. Attracting the news: Algorithms, platforms, and reframing incidental exposure. *Journalism*, 21(8):1067–1082, 2020.

[143] Michael X Delli Carpini and Scott Keeter. *What Americans know about politics and why it matters*. Yale University Press, 1996.

[144] Markus Prior. News vs. entertainment: How increasing media choice widens gaps in political knowledge and turnout. *American Journal of Political Science*, 49(3):577–592, 2005.

[145] Markus Prior. *Post-broadcast democracy: How media choice increases inequality in political involvement and polarizes elections*. Cambridge University Press, 2007.

[146] Jason Barabas and Jennifer Jerit. Estimating the causal effects of media coverage on policy-specific knowledge. *American Journal of Political Science*, 53(1):73–89, 2009.

[147] Jennifer Jerit, Jason Barabas, and Toby Bolsen. Citizens, knowledge, and the information environment. *American Journal of Political Science*, 50(2):266–282, 2006.

[148] Arthur Lupia and Mathew D McCubbins. *The democratic dilemma: Can citizens learn what they need to know?* Cambridge University Press, 1998.

[149] C Mont'Alverne, A Ross Arguedas, S Banerjee, B Toff, R Fletcher, and R Nielsen. The electoral misinformation nexus: How news consumption, platform use, and trust in news influence belief in electoral misinformation. *Public Opinion Quarterly*, 2024.

[150] Sacha Altay, Ewa Hoes, and Magdalena Wojcieszak. News on social media boosts knowledge, belief accuracy, and trust: A field experiment on instagram and WhatsApp, July 30 2024. `https://doi.org/10.31234/osf.io/hq5ru`.

[151] Michael X Delli Carpini. In search of the informed citizen: What Americans know about politics and why it matters. *The Communication Review*, 4(1):129–164, 2000.

[152] Michael X Delli Carpini and S. Keeter. The internet and an informed citizenry, 2002.

[153] Danny Hayes and Jennifer L Lawless. As local news goes, so goes citizen engagement: Media, knowledge, and participation in US house elections. *The Journal of Politics*, 77(2):447–462, 2015.

[154] Pablo Barberá, John T Jost, Jonathan Nagler, Joshua A Tucker, and Richard Bonneau. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological science*, 26(10):1531–1542, 2015.

[155] Richard Fletcher, Craig T Robertson, and Rasmus Kleis Nielsen. How many people live in politically partisan online news echo chambers in different countries? *Journal of Quantitative Description: Digital Media*, 1, 2021.

[156] Alan I Abramowitz and Kyle L Saunders. Is polarization a myth? *The Journal of Politics*, 70(2):542–555, 2008.

[157] Pablo Barberá, Andreu Casas, Jonathan Nagler, Patrick J Egan, Richard Bonneau, John T Jost, and Joshua A Tucker. Who Leads? Who Follows? Measuring Issue Attention and Agenda Setting by Legislators and the Mass Public Using Social Media Data. *American Political Science Review*, 113(4):883–901, 2019.

[158] Yanna Krupnikov and John Barry Ryan. *The Other Divide*. Cambridge University Press, Jan 2022. Google-Books-ID: qnRUEAAAQBAJ.

[159] Morten Skovsgaard and Kim Andersen. Conceptualizing news avoidance: Towards a shared understanding of different causes and potential solutions. *Journalism studies*, 21(4):459–476, 2020.

[160] Mikko Villi, Tali Aharoni, Keren Tenenboim-Weinblatt, Pablo J. Boczkowski, Kaori Hayashi, Eugenia Mitchelstein, Akira Tanaka, and Neta Kligler-Vilenchik. Taking a break from news: A five-nation study of news avoidance in the digital era. *Digital Journalism*, 10(1):148–164, Jan 2022.

[161] Judith Moeller, Claes De Vreese, Frank Esser, and Ruth Kunz. Pathway to political participation: The influence of online and offline news media on internal efficacy and turnout of first-time voters. *American behavioral scientist*, 58(5):689–700, 2014.

[162] Veronika Karnowski, Anna S Kümpel, Larissa Leonhard, and Dominik J Leiner. From incidental news exposure to news engagement. how perceptions of the news post and news usage patterns influence engagement with news articles encountered on Facebook. *Computers in Human Behavior*, 76:42–50, 2017.

[163] Jessica T. Feezell. Agenda setting through social media: The importance of incidental news exposure and social filtering in the digital era. *Political Research Quarterly*, 71(2):482–494, Jun 2018.

[164] J Taylor. Meta says Facebook cannot solve media industry's 'issues' as it defends ending payments for news in australia. *The Guardian*, 2024.

[165] J Taylor. Meta claims news is not an antidote to misinformation on its platforms. *The Guardian*, 2024.

[166] Steve Rathje, Claire Robertson, William J Brady, and Jay J Van Bavel. People think that social media platforms do (but should not) amplify divisive content. *Perspectives on Psychological Science*, page 17456916231190392, 2023.