

An Analysis of the Formation Mechanism of Social Media Users' Opinion Polarization under the Information Cocoon Effect

Zijun Zhang*

Graduate School of Humanities and Social Sciences, The University of Melbourne, Melbourne, Victoria, Australia

Abstract. The phenomenon of opinion polarization on social media has attracted significant scholarly attention. However, existing studies mostly focus on either user behavior or technological factors in isolation, offering limited insight into the deeper, systemic mechanisms behind its emergence. This paper examines the interactive relationship between users' subjective preferences and algorithmic recommendations, analyzing their synergistic role in driving polarization. The findings suggest that opinion polarization is not driven by a single linear factor but rather emerges as a dynamic outcome shaped by the interplay between external technological forces and internal cognitive motivations. Individuals, influenced by psychological mechanisms such as cognitive dissonance and confirmation bias, tend to seek out information aligned with their preexisting positions. Algorithmic recommendations further reinforce this tendency by constructing increasingly homogenous information environments, thereby intensifying echo chambers and filter bubbles. Based on these insights, the paper recommends that platform algorithms be optimized to incorporate mechanisms for diverse information exposure and improved visibility of heterogeneous content. Furthermore, enhancing users' information literacy, fostering rational expression, and cross-group dialogue are proposed as essential strategies to mitigate the adverse effects of opinion polarization on the public opinion ecology.

1 Introduction

With the rapid advancement of internet technologies, social media has become a crucial platform for information dissemination and opinion formation. These platforms utilize personalized algorithmic recommendation systems to provide users with efficient information services, while also placing them in highly homogenized information cocoons [1]. Immersed in such environments, users tend to engage primarily with content that aligns with their existing views and interests, reinforcing their preexisting cognitive frameworks [2]. Under the combined influence of information filtering and selection preference, social media has gradually evolved into a fertile ground for opinion polarization. As public understanding of social issues becomes increasingly divided and antagonistic, efforts to build societal

* Corresponding author: zijuzhang4@student.unimelb.edu.au

consensus face mounting challenges. This phenomenon has sparked growing scholarly interest and public concern across the globe.

Specifically, existing studies on the causes of opinion polarization generally follow two main theoretical approaches. The first emphasizes the structural influence of technological algorithms, drawing on the concepts of the filter bubble and echo chamber to highlight how algorithms constrain users' exposure to diverse information [3]. The second centers on selective exposure theory, which focuses on individual psychological and behavioral tendencies in actively filtering information [2]. However, both perspectives are often limited in scope. While some fall into the trap of technological determinism, attributing polarization solely to platform mechanisms, others overstate user agency and neglect the platform's structural role. To advance the field, it is essential to move beyond such singular perspectives and examine polarization as a product of the dynamic interaction between technological infrastructures and user behavior.

In response to these theoretical limitations, this paper adopts an integrated analytical perspective to examine the dynamic formation of opinion polarization in social media environments. Emphasizing the interplay between individual-level cognitive mechanisms and platform-level technological structures, the paper draws on multiple theoretical frameworks—including selective exposure theory, the filter bubble hypothesis, and media ecology. Through this multidimensional lens, it seeks to offer a more systematic explanation of the generative logic of polarization and to identify potential avenues for effective intervention.

2 The concept, characterization, and impact of opinion polarization

2.1 Basic concepts and theoretical sources

Opinion polarization refers to the phenomenon for public attitudes on specific issues become increasingly extreme. It is typically characterized by growing intra-group consensus and intensified inter-group antagonism [4]. The concept originated in the social psychology field. In the 1960s, social psychologists Moscovici and Zavalloni introduced the notions of "polarization" and the "polarization effect" through experimental studies. They found that group discussions often lead individuals to adopt more extreme positions consistent with their initial attitudes [5]. This foundational insight paved the way for subsequent research on group decision-making, social cognition, and the dynamics of public opinion.

Sunstein further advanced the theoretical framework of opinion polarization within the context of online communication. He argued that when individuals are exposed exclusively to information that reinforces their preexisting beliefs, repeated confirmation drives their attitudes toward greater extremity [6]. His work marked a conceptual shift from group-level psychological processes to structural features of media environments, offering a critical theoretical lens for understanding polarization in the era of social media.

Today, opinion polarization is recognized as a complex social phenomenon shaped by both macro-level structural factors and micro-level psychological mechanisms, and has gradually evolved into a multidisciplinary theoretical construct. With the rise of social media and algorithmic recommendation systems, polarization has become a key concept for analyzing current transformations in public discourse, digital governance, and societal fragmentation.

2.2 Representations of opinion polarization in social media

Building on existing research, opinion polarization typically manifests in three forms. The first is attitudinal polarization, in which individuals, when exposed to information that confirms their prior beliefs, tend to further reinforce those beliefs, resulting in increasingly extreme positions [7]. The second is interactional polarization, referring to users' tendency to engage with like-minded individuals, leading to the formation of highly homogeneous in-group networks and segregated out-group boundaries. This interactional structure deepens intra-group consensus and intensifies hostility and opposition toward out-groups, thereby heightening communication barriers and, in some cases, contributing to group polarization [8]. The third form is affective polarization, characterized by heightened emotional identification within groups and antagonism toward those outside the group. In online public spheres, polarization often extends beyond ideological divergence to include overt expressions of negative emotions toward opposing groups. These sentiments are often expressed through abusive language, emotionally charged commentary, and labeled language with mocking connotations. Amplified by algorithmic recommendation systems, these affective expressions can spread rapidly across platforms, generating broad emotional resonance and fueling collective mobilization [9].

2.3 Multidimensional impact of opinion polarization

The impact of opinion polarization is multifaceted. At the individual level, it fosters cognitive closure and reinforces preexisting biases. Prolonged immersion in highly homogeneous information environments facilitates the development of narrow and increasingly extreme cognitive frameworks. This pattern of thinking can lead to mental rigidity, ultimately weakening individuals' capacity for comprehensive understanding and rational decision-making [10].

In society, polarization undermines the formation of social consensus and contributes to the fragmentation of public discourse. As emotionally charged expressions come to dominate the public sphere, meaningful dialogue and negotiation between opposing groups become increasingly difficult. This trend not only threatens social cohesion but also hampers democratic deliberation and the effectiveness of public governance. In more severe cases, the persistent spread of polarization may deepen social divisions and escalate group-based conflicts, posing risks to social solidarity and stability.

At the level of the information ecosystem, polarization profoundly shapes the production and circulation of content. In pursuit of higher user engagement and platform retention, algorithmic systems may preferentially promote sensationalist or extreme content. Such content often features emotionally driven rhetoric, selective presentation of facts, or even misinformation, leading to the systemic degradation of information quality on digital platforms [10]. This will not only damage the health of the information ecosystem but also weaken the public's trust in media platforms.

Overall, opinion polarization poses significant risks not only to individual cognition and psychological health but also to public consensus and the information ecosystem. These risks emphasize the urgent need to analyze polarization through a systemic lens and to develop targeted intervention strategies and governance mechanisms aimed at mitigating its harmful effects.

3 Mechanisms of opinion polarization

3.1 Endogenous dynamics of individual choice

The formation of opinion polarization is largely driven by the combined influence of individuals' subjective preferences and underlying psychological mechanisms during the process of information selection. One of the most prominent theoretical explanations is the selective exposure theory. It posits that individuals tend to seek out information consistent with their existing views and attitudes, while avoiding content that may trigger cognitive dissonance [11]. This tendency reflects a fundamental aspect of human behavior—the pursuit of psychological equilibrium and cognitive consistency through selective engagement with agreeable information.

From a psychological perspective, selective exposure is primarily shaped by two cognitive mechanisms: cognitive dissonance and confirmation bias [12]. These mechanisms create a self-reinforcing loop, whereby individuals repeatedly consume information aligned with their preexisting views. Such repetitive exposure strengthens existing cognitive frameworks and value orientations, ultimately contributing to the radicalization of attitudes and ideological entrenchment.

On algorithm-driven social media platforms, the mechanism of selective exposure becomes more pronounced. These platforms offer users considerable autonomy in filtering information, enabling them to construct highly personalized information environments based on their interests, ideological leanings, and cognitive needs [12]. Through following, liking, commenting, and sharing, users continuously provide feedback that reflects their preferences, which in turn informs algorithmic content recommendations. This recursive process reinforces users' existing informational biases, intensifies content homogeneity, and increases the likelihood of individuals becoming enclosed within information cocoons and echo chambers. Together, these dynamics contribute to the emergence and amplification of opinion polarization.

Nonetheless, the applicability of selective exposure theory in digital media contexts remains contested. Some scholars argue that while users generally prefer information congruent with their views, this does not imply a complete avoidance of opposing perspectives [13]. Information selection is shaped not only by personal preference but also by external factors, including media context and social relationships.

The selective exposure theory provides a critical cognitive foundation for understanding opinion polarization. However, neglecting its coupling with technological structures and social interaction mechanisms makes it difficult to fully understand the complex generation logic of the current polarization phenomenon. In the context of digital communication, selective exposure should not be treated as an isolated psychological tendency, but rather examined within the broader structural framework of user behavior and algorithmic systems.

3.2 External dynamics of technology architecture

In social media environments, the formation of opinion polarization is shaped not only by users' subjective choices but also by the influence of platform technologies. In particular, personalized algorithmic recommendation systems have emerged as a key external force that shapes users' information contact path and public opinion ecosystem.

These recommendation mechanisms precisely control information flows by constructing a multi-layered information visibility structure. Mainstream platforms currently use technologies such as hotspot algorithms, content-based collaborative filtering algorithms, and user-based collaborative filtering algorithms, which play roles at the macro (overall public opinion atmosphere), meso (circle public opinion atmosphere), and micro (individual opinion

atmosphere) levels [8]. By dynamically adjusting content based on users' clicking, browsing, and interaction behaviors, algorithms construct highly customized information spaces that fits users' interests and ideological positions.

Driven by the logic of the attention economy, platforms continuously optimize their recommendation models to maximize user engagement and platform revenue [1]. While such mechanisms enhance the efficiency of information acquisition, they also exacerbate informational filtering. As users are repeatedly exposed to content aligned with their preferences, the diversity of information they encounter narrows, leading to increasingly homogenous content environments. This process reduces exposure to dissenting perspectives, reinforces cognitive closure, and intensifies the polarization of attitudes.

This "stratosphere" information architecture, shaped by platform algorithms, serves as the technological foundation for the emergence of echo chambers and filter bubbles. It makes social media platforms an amplifier of polarization and a manipulator of public opinion. Therefore, understanding the structural impact of algorithmic recommendation systems and their role in shaping opinion polarization is essential for explaining the mechanisms underlying polarization in the current digital media landscape.

4 Discussion

4.1 The logic of interactivity in polarization pathways

The polarization of social media users' opinions is not driven by a single factor, but rather emerges through the long-term, dynamic interplay between individual behavior and platform technologies. Within the highly personalized and interactive architecture of social media platforms, users' information exposure and feedback behaviors form a self-reinforcing system. On the one hand, users engage in selective exposure based on existing cognitive structures, emotional stances, and social identity preferences to maintain cognitive consistency, emotional stability, and identity coherence. On the other hand, platform algorithms continuously record and analyze user behavior to optimize content delivery, further reinforcing users' inherent preferences and cognitions [1]. In this way, the logic of platform technology is dynamically embedded in user behavior, jointly shaping the informational environment in which users operate.

This "behavior-algorithm-behavior" circular path highlights the non-linear and evolving nature of the polarization process. Therefore, the underlying logic of opinion polarization is the outcome of multidimensional variables interacting and co-evolving, rather than being linearly determined by a single factor.

4.2 Theory deepening and integration

Based on the above analysis, this paper argues that the formation of opinion polarization in social media environments should be understood as a systemic process shaped by the interplay of multiple factors, encompassing cognitive and psychological mechanisms, platform technological architecture, and social network interactions.

To address the current theoretical limitations, this study proposes an integrated analytical perspective. On one hand, this perspective acknowledges the structural influence of algorithmic recommendation systems, social configurations, and dissemination mechanisms in shaping public opinion and reinforcing ideological positions. On the other hand, it also emphasizes users' agency, including their strategic behaviors in content selection, emotional expression, and group interaction. The core of this approach is the recognition that technological structures and behavioral logics function as dynamic variables in a reciprocal

relationship, rather than as isolated or static components. The relationship between platforms and the individual is not a static one of tool and user, but a co-constructive and negotiated network formed through ongoing interaction and mutual adaptation.

Within this framework, opinion polarization can be re-understood as a co-produced outcome of user--platform interaction within a jointly shaped information ecosystem. Specifically, individuals' selective exposure constitutes the cognitive precondition for polarization. The platform's information filtering and content recommendation mechanisms enhance the visibility of homogeneous content, structurally reinforcing the information cocoon effect. Users engage in affective mobilization based on group affiliation, value alignment, and emotional resonance, which further radicalizes opinion expression [8]. Moreover, the highly interactive features and community aggregation mechanisms of social media platforms institutionalize these cognitive, affective, and structural forces into routine information dissemination processes, thereby continuously amplifying the diffusion of polarized opinions.

These elements do not operate in a linear sequence but function as interdependent variables within a dynamic system of interaction, collectively forming the systemic logic through which opinion polarization evolves. Ultimately, this multidimensional and co-evolutionary theoretical perspective offers a more comprehensive lens through which to overcome the limitations of conventional single-cause or single-mechanism explanations, thereby enabling a more integrative understanding of polarization in the digital media age.

5 Conclusion

This paper offers a systematic analysis of the formation mechanisms of opinion polarization in social media environments, arguing that such polarization is not the product of a single factor, but rather the outcome of the interaction between users' subjective choices and algorithmic recommendation systems. On one hand, platforms push customized content based on users' historical behaviors and preferences, thereby reinforcing the homogeneity of information exposure. On the other hand, users engage in selective exposure, actively receiving content that conforms to their existing views, which further strengthens algorithm's biased push. This reciprocal process creates a self-reinforcing cycle in which attitudinal, interactional, and affective polarization become intertwined, shaping the multilayered landscape of contemporary polarization on social media.

It highlights the need to move beyond single-perspective explanations and instead focus on the coupling mechanisms between technological systems and human agency. Future research could further explore how different platform architectures shape varying expressions of polarization, as well as the role of user-led interventions in disrupting information cocoons. Such inquiries may offer valuable theoretical and practical insights for fostering a more open, pluralistic, and healthy public discourse ecosystem.

References

1. Y. Z. Liu, Q. S. Hu, Social reflection based on algorithm recommendation mode: Individual dilemma, group polarization and media publicity. *Media Econ. Manag. Res.* **1**, 192-212 (2021)
2. N. J. Stroud, Polarization and partisan selective exposure. *J. Commun.* **60(3)**, 556-576 (2010)
3. E. Pariser, *The filter bubble: What the Internet is hiding from you*. Penguin UK (2011)

4. J. Lee, Y. Choi, Effects of network heterogeneity on social media on opinion polarization among South Koreans: Focusing on fear and political orientation. *Int. Commun. Gaz.* **82(2)**, 119-139 (2020)
5. S. Moscovici, M. Zavalloni, The group as a polarizer of attitudes. *J. Pers. Soc. Psychol.* **12(2)**, 125 (1969)
6. C. R. Sunstein, *Republic.com*. Princeton University Press (2001)
7. M. Wojcieszak, 'Don't talk to me': Effects of ideologically homogeneous online groups and politically dissimilar offline ties on extremism. *New Media Soc.* **12(4)**, 637-655 (2010)
8. Q. H. Yan, The mechanism of algorithmic recommendations' influence on netizen polarization: A multi-level analysis from the perspective of public opinion visibility. *Mod. Commun. (J. Commun. Univ. China)* **8**, 12-25 (2024)
9. S. Rathje, J. J. Van Bavel, S. Van Der Linden, Out-group animosity drives engagement on social media. *Proc. Natl. Acad. Sci. U.S.A.* **118(26)**, e2024292118 (2021)
10. H. Y. Zhang, Y. N. Ding, S. Q. Guo, Research on the mechanism of user group polarization under the information cocoon effect. *J. Inf.* **3**, 132-144 (2024)
11. W. Hart, et al., Feeling validated versus being correct: a meta-analysis of selective exposure to information. *Psychol. Bull.* **135(4)**, 555 (2009)
12. J. Shen, An empirical study of social media factors influencing youth opinion polarization. *J. Shanghai Jiaotong Univ. (Philos. Soc. Sci.)* **32(6)**, 133-150 (2024)
13. R. K. Garrett, Echo chambers online? Politically motivated selective exposure among Internet news users. *J. Comput. Mediat. Commun.* **14(2)**, 265-285 (2009)