Proceedings of the 7th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/2025.26216

# *The Formation Mechanism of Social Media Public Opinion Polarization under Algorithmic Bias*

**Yixin He**

*College of Humanities and Social Sciences, The Chinese University of Hong Kong, Shenzhen, China*
*h222035012@163.com*

***Abstract.*** In the era of Web 3.0, data-driven recommendation systems dominate the dissemination of social media information, leading to issues such as cognitive imbalance, public opinion polarization, group opposition, and hidden risks of social fragmentation. This study reveals the mechanism of algorithmic bias on public opinion polarization, providing reference for understanding the social media public opinion ecology.

***Keywords:*** Algorithmic bias, Social media, Polarization of public opinion, User behavior

## 1. Introduction

In the profound changes of the Web 3.0 era, data-driven recommendation systems have become the core engine for social media information dissemination. While reshaping the way the public obtains information, they have also quietly caused serious problems of cognitive imbalance [1]. The evolution trend of public opinion is increasingly dominated by algorithms [2]. In this context, the phenomenon of polarization of public opinion is becoming increasingly prominent [3], which not only erodes the public space for rational discussion, but also harbors deep-seated concerns of shaking social consensus and causing social tearing [4]. Algorithmic bias is not a simple concept with a single dimension, but rather a hidden value orientation and uncertainty in the construction of algorithmic models [5-6]. Polarization of public opinion is a dynamic process in which different social groups gradually move towards opposing extremes in terms of viewpoints, emotions, and identity recognition [7-8]. As a public opinion arena, social media shapes the generation and evolution trajectory of public opinion [9].

In the research context of algorithmic bias, there has been a paradigm shift from technology neutrality to value loading. Early studies often viewed algorithms as purely technical tools [10], but as research deepened, scholars gradually realized that algorithms embedded developers' value judgments [11]. In the study of the formation mechanism of public opinion polarization, theories such as filter bubbles have revealed the impact of information environment on group differentiation, but there are also certain limitations [12]. The polarization amplification effect of existing models, such as collaborative filtering algorithms, has been partially empirically validated, but there is still room for further exploration of the systematic and complex mechanisms [13].

Based on this, this article focuses on the impact of algorithmic bias on social media public opinion polarization. By constructing a multidimensional analysis model, it reveals the significant positive impact and threshold effect of algorithmic bias on public opinion polarization, and clarifies

DOI: 10.54254/2755-2721/2025.26216

the differentiated pathways of algorithmic bias in different dimensions. In theory, this article deepens the understanding of the complex relationship between algorithmic bias and public opinion polarization, providing a systematic analytical framework and empirical basis for related research; In practice, this article helps to understand the evolutionary laws of social media public opinion ecology, providing targeted reference directions for alleviating public opinion polarization, maintaining rational public space, and social stability.

## 2. Research design

### 2.1. Research framework construction

When deconstructing the complex relationship between algorithmic bias and public opinion polarization, a single disciplinary perspective can easily fall into methodological traps. This article is based on the structured theory of systems science and communication studies, and constructs a multidimensional analysis model with the following expression:

$$Polarization_t = f(SNA(Net_{t-1}), CC(Algo_t, User_t), \varepsilon_t) \tag{1}$$

Among them, SNA (·) is a social network analysis function; CC (·) is a computational propagation function; $\varepsilon_t$ is the system noise term.

The core logic lies in the dynamic evolution process of public opinion polarization, which is the independent variable of algorithm bias, the mediating variable of user cognitive behavior, and the moderating variable of social network structure, through a three dimensional deconstruction mechanism.

Algorithmic bias has a nonlinear amplification effect on user cognition. Algorithmic bias implants initial bias seeds into the user cognitive system through the interaction between Selective Exposure and Cognitive Heuristics. The dual filtering expression for information filtering is as follows:

$$Pr\left(Exposure \middle| Algo\right) = \frac{e^{\alpha \cdot Similar + \beta \cdot Emotional}}{e^{\gamma \cdot Diversity}} \tag{2}$$

Among them, $S_{similar}$ is the weight of content similarity, $E_{emotional}$ is the intensity of emotional arousal, and $D_{diversity}$ is the diversity penalty factor (α, β, γ are platform preset parameters)

Based on the attention based bounded rationality model, the processing depth δ c of biased content by users shows a marginal increasing effect, expressed as follows:

$$\delta_c = \lambda \cdot log(1 + \omega \cdot \mathrm{Bias}_{\mathrm{intensity}}) \tag{3}$$

Where λ is the cognitive elasticity coefficient.

### 2.2. Data collection and processing

This study constructs a three-stage funnel-shaped data collection system, ensuring sample representativeness through cross platform complementary strategies. The specific dataset collected is shown in Table 1.

DOI: 10.54254/2755-2721/2025.26216

Table 1. Dataset

| Data Layer | Collection Platform/Method | Time and Space Scope | Sample Size |
|---|---|---|---|
| Core Behavioral Data | Twitter API v2 (Academic-level Access) | January 2023 - December 2023, Sino-US hot social issues | 4.2 million original tweets |
| | Weibo Super Topic Crawler (Python Scrapy) | March 2023 - February 2024, 20 controversial topics | 2.8 million blog posts/comments |
| Algorithm Output Data | Self-developed browser plugin WebTracker | Installed by 2,000 volunteers to track recommendation streams | 6.1 million pushed contents |
| User Cognition Data | Stratified Sampling Questionnaire (LimeSurvey) | 500 users each from China, Britain and America, cognitive flexibility test | 1,382 valid questionnaires |

## 2.3. Variable definition

The main variables and definitions involved in this article are shown in Table 2.

Table 2. Variable definition

| Variable Category | Variable Name | Symbol | Data Source |
|---|---|---|---|
| Independent Variable | Comprehensive Index of Algorithm Bias | AlgoBias | Recommendation Stream Crawling Data |
| | Data-Level Bias | Rb | WebTracker Plugin Logs |
| | Model-Level Bias | Sd | NLP Analysis of Pushed Content |
| | Feedback-Level Bias | Pr | User Browsing History |
| Dependent Variable | Content Position Polarization | Ediv | Analysis of Tweets/Blog Posts |
| | Network Structure Polarization | ΔQ | Social Network Topology Analysis |
| | Emotional Distribution Polarization | Bc | Comment Sentiment Annotation |
| Mediating Variable | Cognitive Narrowing Index | HI | Analysis of User Browsing History |
| Control Variable | User Activity | Activity | Platform Behavior Logs |
| | Topic Popularity | Heat | Google Trends/Platform Trending Searches |
| | Time Decay Factor | λ | Time-Series Network Fitting |

## 2.4. Model construction

The benchmark model expression is as follows:

$$Polarization_{it} = \beta_0 + \beta_1 AlgoBias_{it} + \beta_2\left(AlgoBias_{it} \times HI_{it}\right) + \sum_{k=1}^{K_1} \gamma_k Control_{kit} + \mu_i + \lambda_t + \varepsilon_{it} \quad (4)$$

The expression of the nonlinear extended model is as follows:

$$\begin{aligned} Polarization_{it} &= \beta_0 + \beta_1 AlgoBias_{it} \cdot \mathbb{I}(AlgoBias_{it} \le \theta) \\ &+ \beta_2 AlgoBias_{it} \cdot \mathbb{I}(AlgoBias_{it} > \theta) \\ &+ \beta_3 NetworkModularity_{it} \times CognitiveRigidity_{it} \\ &+ \mu_i + \lambda_t + \varepsilon_{it} \end{aligned} \quad (5)$$

Among them, i is the user/community unit, and t is the weekly time slice (t=1,2,..., 52); $\mu$ i is the individual fixed effect; $\lambda$ t is the time fixed effect; $\theta$ is the critical value of algorithmic bias intensity.

# 3. Result

## 3.1. Main effect

The main effect results are shown in Table 3, and Model 4 indicates that cognitive narrowing reinforces the effect of algorithmic bias. Model 5 reveals a threshold effect, where the AlgoBias effect is not significant in the low bias area, but significantly increases in the high bias area, and HI and interaction terms remain significant. This indicates that there is a critical value for the impact of algorithmic bias, and the effect amplifies sharply after exceeding it. The moderating effect of cognitive factors remains stable.

Table 3. Main effect test results

| Variable | Model 4 | Model 5 | Method Description |
|---|---|---|---|
| AlgoBias | 0.327*(0.021) | - | Mixed-effects panel model |
| - Low Bias Zone | - | 0.108(0.094) | Hansen threshold regression |
| - High Bias Zone | - | 0.602*(0.137) | Bootstrap iteration 500 times |
| HI | 0.195**(0.088) | 0.211**(0.097) | Control individual/time fixed effects |
| AlgoBias×HI | 0.371**(0.152) | 0.392**(0.173) | Moderating effect test |

## 3.2. Mechanism verification

The results of the mechanism test are shown in Table 4, and all three have passed statistical tests, confirming the differential mechanism of different dimensions of algorithmic bias on public opinion polarization.

Table 4. Bias Component Mechanism Test

| Component Symbol | Action Path | Statistical Evidence |
|---|---|---|
| Rb | Group Representation Bias → Position Opposition | $\Delta$Ediv = 0.73Rb*(0.18) |
| Sd | Semantic Shift → Emotional Polarization | SentExt = 1.88Sd$^2$*(0.79) - 2.31Sd(1.02) |
| Pr | Preference Reinforcement → Information Narrowing | r(Pr,HI) = 0.49*(0.07) |

## 3.3. Sub dimensional evolution

The sub dimensional evolution results are shown in Table 5, where Ediv increased from 1.62 to 2.37, an increase of 46.3%. The Wilcoxon test is significant, indicating a significant increase in the degree of polarization of content stance; The increase in $\Delta$ Q reached 162.5%, and the SAOM model showed highly statistically significant changes, reflecting the severe polarization of the network structure; Bc increased from 0.31 to 0.59, an increase of 90.3%, and the K-S test was significant, indicating a significant improvement in the polarization of emotional distribution.

DOI: 10.54254/2755-2721/2025.26216

Table 5. Evolution characteristics of polarization sub dimensions

| Indicator Symbol | Initial Observation Value | Final Observation Value | Change | Statistical Test |
|---|---|---|---|---|
| Ediv | 1.62 | 2.37 | +46.3% | Wilcoxon Z=18.4*(0.00) |
| ΔQ | 0.08 | 0.21 | +162.5% | SAOM β=1.38**(0.61) |
| Bc | 0.31 | 0.59 | +90.3% | K-S Test D=0.38*(0.000) |

As shown in Figure 1, all three have passed strict statistical tests, confirming the comprehensive and significant intensification of public opinion polarization in terms of content, structure, and emotional dimensions.

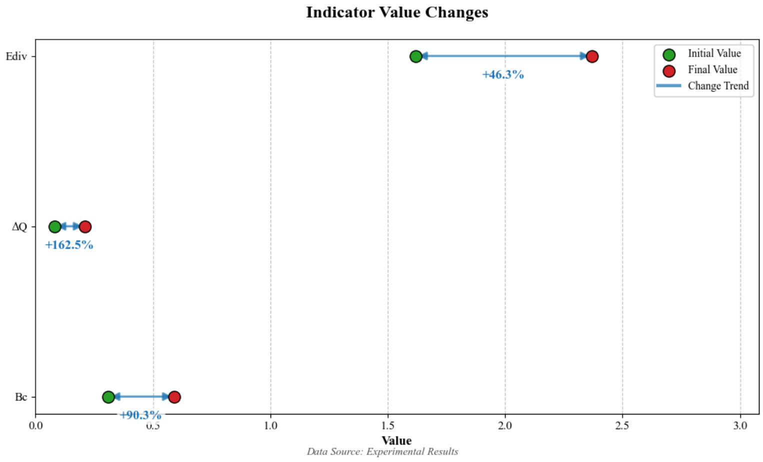


Figure 1. Evolution characteristics of polarization sub dimensions

## 4. Conclusion

This study found that algorithmic bias has a significant positive impact on social media public opinion polarization, and there is a threshold effect. The impact of high bias areas is much stronger than that of low bias areas, and cognitive narrowing will strengthen this effect; Data level, model level, and feedback level biases act on polarization through different paths, with content stance, network structure, and emotional distribution polarization significantly exacerbated, with network structure polarization showing the greatest increase; The cross platform robustness test confirms the reliability of the above conclusion and reveals the systematic mechanism of algorithmic bias catalyzing public opinion polarization.

## References

[1] Zhang, H., & Lu, Y. (2025). Web 3.0: Applications, Opportunities and Challenges in the Next Internet Generation. Systems Research and Behavioral Science.

[2] Sætra, H. S. (2019). The tyranny of perceived opinion: Freedom and information in the era of big data. Technology in Society, 59, 101155.

[3] Yi, W. (2023). An Analysis of the Information Cocoon Effect of News Clients: Today's Headlines as an Example. The Frontiers of Society, Science and Technology, 5(9).

[4] Liu, W., & Zhou, W. (2022). Research on solving path of negative effect of “information cocoon room” in emergency. Discrete Dynamics in Nature and Society, 2022(1), 1326579.

[5] Kordzadeh, N., & Ghasemaghaei, M. (2022). Algorithmic bias: review, synthesis, and future research directions. European Journal of Information Systems, 31(3), 388-409.

DOI: 10.54254/2755-2721/2025.26216

[6] Panch, T., Mattie, H., & Atun, R. (2019). Artificial intelligence and algorithmic bias: implications for health systems. Journal of global health, 9(2), 020318.
[7] Strickler, R. (2018). Deliberate with the enemy? Polarization, social identity, and attitudes toward disagreement. Political Research Quarterly, 71(1), 3-18.
[8] Iyengar, S., Sood, G., & Lelkes, Y. (2012). Affect, not ideology: A social identity perspective on polarization. Public opinion quarterly, 76(3), 405-431.
[9] Matakos, A., Terzi, E., & Tsaparas, P. (2017). Measuring and moderating opinion polarization in social networks. Data Mining and Knowledge Discovery, 31(5), 1480-1505.
[10] Johnson, G. M. (2021). Algorithmic bias: on the implicit biases of social technology. Synthese, 198(10), 9941-9961.
[11] Draude, C., Klumbyte, G., Lücking, P., & Treusch, P. (2020). Situated algorithms: a sociotechnical systemic approach to bias. Online Information Review, 44(2), 325-342.
[12] Zimmer, F., Scheibe, K., Stock, M., & Stock, W. G. (2019, January). Echo chambers and filter bubbles of fake news in social media. Man-made or produced by algorithms. In 8th annual arts, humanities, social sciences & education conference (pp. 1-22).
[13] Stinson, C. (2022). Algorithms are not neutral: Bias in collaborative filtering. AI and Ethics, 2(4), 763-770.