Proceedings of ICILLP 2025 Symposium: Property Law and Blockchain Applications in International Law and Legal Policy
DOI: 10.54254/2753-7048/2025.LD29977

# *Can AI Trolls Polarize Public Opinion? — A Case Study Based on the 2024 Jiangsu Sihong "AI Troll" Incident*

**Jiamei Wu**

*School of Economics and Management, Communication University of China, Beijing, China*
*wujiameimm@163.com*

***Abstract.*** Generative AI is changing the cost structure of information production and dissemination. "AI-enabled astroturfing," which leverages AI to generate content at low cost and scale, poses a potential threat to the online public opinion ecosystem. This article explores whether and how AI astroturfing exacerbates public opinion polarization. Using a case study approach, the study delves into the August 2024 incident involving Zong Mou and others in Sihong, Jiangsu Province, who manipulated public opinion to boost user traffic. The analysis is conducted from the perspective of the platform economy and information costs. While generative AI was not directly employed in this case, its operational model reveals the infrastructure and economic logic of the astroturfing industry. The study finds that AI astroturfing can leverage this infrastructure to exacerbate polarization through two mechanisms: first, through the large-scale production of content, artificial majority opinions are created, leading to apparent polarization; second, through interaction with algorithmic recommendation systems, AI-generated customized content can reinforce filter bubbles and echo chambers, exacerbating polarization. The conclusion suggests that governance of AI astroturfing should go beyond simple content moderation and move toward a comprehensive framework that holds platforms accountable for economic responsibilities, enhances algorithmic transparency, and strengthens digital literacy education.

***Keywords:*** AI water army, Public opinion polarization, Generative AI, Jiangsu Sihong case, Social governance

## 1. Introduction

While generative AI improves content production efficiency, it also provides new technological possibilities for public opinion manipulation. Traditional "online water armies" are limited by high labor costs and low-quality content homogeneity; generative AI significantly reduces the marginal cost of content production, making automated, scalable, and personalized false information generation possible [1]. This technological change may exacerbate the spread of false information, thereby posing a significant challenge to the online public opinion ecosystem. Moreover, opinion polarization has emerged as a pressing social issue, characterized by the public's rigid attitudes and emotional opposition on specific issues, which hinders the formation of social consensus [2].

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-7048/2025.LD29977

Existing research often links this phenomenon to filter bubbles and “echo chambers” caused by algorithmic recommendation systems [3-4].

This study therefore raises the core question: Do AI-enabled online troll armies (hereinafter referred to as AI trolls) exacerbate public opinion polarization, and what are the underlying mechanisms? Answering this question requires empirical case analysis. The August 2024 verdict in the Jiangsu Sihong case involving Zong Mou and others for manipulation of user traffic and comments provides a case example. This case is a prime example of the "Clean Net 2024" campaign. Zong Mou's studio controlled over 6,000 cloud phones, using group control software and scripts. Its scale and technology demonstrate the industrialized nature of the online troll industry. Although this case did not directly utilize generative AI, its automated, scalable infrastructure provides a foundation for studying the deployment capabilities of AI trolls.

This article, using a literature review and case study approach, begins from the perspective of media economics. It first analyzes the economic logic and operational model of the Sihong case, then deduces the potential mechanisms by which this model, combined with AI technology, influences public opinion polarization. Finally, it offers corresponding governance implications. This research provides theoretical basis and practical guidance for relevant policy formulation and public opinion governance.

## 2. Literature review

Academically, “Internet water army” is defined as “astroturfing”, which means “organized activities designed to imitate grassroots behavior” [5]. Traditional “water army” research mainly focuses on technical identification. The emergence of generative AI is changing the research agenda in this field. “AI water army” (AI-enabled astroturfing) is no longer limited to zombie accounts or repetitive comments, but can generate grammatically fluent, seemingly real, and even deep fake graphic content [1]. From the perspective of media economics, the core impact of AI technology is that it greatly reduces the “marginal cost” of producing false information [6]. Under the traditional model, organized manipulation of public opinion requires high manpower or technical investment. The low-cost nature of AI enables small organizations and even individuals to have the potential to initiate large-scale public opinion manipulation. This is not only an increase in quantity, but may also bring about a change in nature. Opinion polarization is another core variable of this study. The study of its causes has shifted from the early social structure to the digital technology environment [7]. In the digital age, the “black box” characteristics of platform algorithms have attracted much attention. Pariser's filter bubble theory and Sunstein's echo chamber theory both point out that the strengthening of selective exposure - that is, algorithms continuously push information that users prefer - may lead to a narrow perspective of users and reduce tolerance for dissenting views. In addition, psychological mechanisms such as hostile media perception have also been confirmed. Individuals with strong partisanship are more likely to view neutral reports as biased towards the opposing side, which solidifies distrust and opposition [8].

The relationship between AI and public opinion polarization is a subject of ongoing debate in existing research. There are two predominant perspectives. On the one hand, AI is believed to be likely to exacerbate polarization. Political economic models point out that AI-driven recommendation algorithms and targeted advertising will systematically exacerbate group divisions [9]. AI reduces the production and dissemination costs of propaganda and misinformation, making it more widespread and more targeted, which may exacerbate polarization [10]. On the other hand, some experimental studies have reached the opposite preliminary conclusion. For example, Bai et al. found that when information was clearly labeled as "AI-generated", subjects believed it was less

DOI: 10.54254/2753-7048/2025.LD29977

biased, which to some extent reduced hostility towards opposing views [11]. This theoretical inconsistency highlights the entry point of this study. The specific impact of AI on polarization may be highly dependent on how it is used—whether as an overt auxiliary tool or a covert manipulation tool. The Jiangsu Sihong case, as a real-world example, provides a sample for observing the infrastructure and economic logic of "troll army" campaigns, helping to deduce the potential impact of AI in covert scenarios.

## 3. Case study: The 2024 Jiangsu Sihong "AI water army" case

### 3.1. Case overview

On August 9, 2024, the Sihong County People's Court in Suqian City, Jiangsu Province, issued a first-instance judgment in the illegal business operation case of Zong and others. This case is one of the typical cases publicized as part of the Ministry of Public Security's "Clean Net 2024" special campaign. The facts of the case are as follows: To boost the influence of the entertainment marketing account he operated, the principal offender, Ji, instructed Zong to establish a studio. This studio utilized over 6,000 cloud phones and exercised centralized control through group control software and scripts. Its primary activity was to provide paid "inflated comment manipulation" services on social platforms such as Weibo to fan groups and the livestreaming industry, specifically bulk false forwarding, likes, and comments. The gang reaped tens of millions of yuan in illegal profits, and the principal offender, Ji, was ultimately sentenced to fixed-term imprisonment and fined for illegal business operations.

The public coverage of this case has heightened public awareness of the industry chain of fabricated online data and traffic. Authoritative media coverage has exposed the operating methods of this shady industry to the public eye and deepened public doubts about the authenticity of trending searches and viral content.

### 3.2. Case model analysis: infrastructure and economic logic

Analysis of the Sihong case reveals that the tools used were group control software and scripts, rather than generative AI technology. However, this article argues that the operational model demonstrated in this case provides a crucial blueprint for understanding the potential deployment of AI trolls. First, the case reveals the industrialized infrastructure of troll army operations. Over 6,000 cloud phones formed a vast account matrix, with group control software acting as a command system. This model enabled public opinion manipulation to move beyond the inefficient manual operations of earlier eras, achieving automation and scale. This infrastructure can be considered a potential tool for future deployment of AI trolls. Second, the conviction in this case—"illegal business operations"—clarifies the economic logic behind the operation. The group's merchandise was fabricated internet traffic. In the platform economy, high-traffic data, such as high likes and comments, can influence the platform's recommendation algorithms, garnering increased exposure and ultimately monetization through advertising or livestreaming. The group's criminal manipulation of public opinion exemplifies this profit-seeking model based on traffic manipulation.

### 3.3. Impact mechanism deduction: potential changes from "script" to "AI"

The "scripted online commentary" model used in the Sihong case provides a foundation for deducing the impact of generative AI intervention. When AI technology is integrated into this type of infrastructure, two changes are expected.

DOI: 10.54254/2753-7048/2025.LD29977

First, the integration of AI technology would lead to intelligent content production. The scripts used in the Sihong case presumably generated content consisting primarily of highly repetitive, low-information text such as "support" and "agree." If these scripts were replaced with generative AI, the infrastructure would be capable of rapidly producing thousands of comments with diverse content and natural tones. It could even mimic the tone of specific groups. The fidelity and deceptiveness of this content far surpass those of traditional scripts, significantly increasing the difficulty for platforms and users to identify and identify it.

Second, the introduction of AI technology would enable a precise polarization mechanism. The services used in the Sihong case (fan clubs and livestreaming) are already common areas for emotional online expression. Empowered by AI technology, such services can be applied to more complex social issues. For example, during public events, AI online commentary can be configured with different positions and deployed algorithmically to different discussion forums. They mass-produce, like, and share extreme speech that reinforces bias and reduces the space for rational discussion. By imitating the extreme emotions of real users, AI water armies may induce other users to adopt more opposing expressions, thereby accelerating the polarization of public opinion on the issue.

## 4. Discussion and inspiration: seeking governance in the algorithmic age

Based on case analysis and mechanistic deduction, the industrialized water army infrastructure exemplified by the Sihong case, when combined with AI technology, has significant potential to drive public opinion polarization. This requires corresponding adjustments to social governance. Traditional content moderation, such as post deletion, will be significantly less effective when faced with the large-scale, intelligent content generation driven by AI. A shift in governance is necessary. The conviction in the Sihong case for illegal business operations provides a key lesson: the focus should shift from content to behavior and economics. Crackdowns should focus on manipulation of user traffic and reviews, as well as the shady and gray industries that support it, such as group control software and cloud phones. Platforms should proactively identify "coordinated inauthentic behavior" and, through technical and legal means, increase the costs and risks of such illegal operations.

Platform algorithms play a key role in the process of AI water armies influencing public opinion. AI water armies exploit algorithmic recommendation mechanisms by creating false hype. These algorithms can also unconsciously amplify false or extreme content, creating a negative cycle. Therefore, platform accountability and algorithmic transparency are core governance priorities. Platforms must be encouraged to enhance the transparency of their recommendation logic, introduce necessary third-party audits, and assess the potential impact of their algorithms on polarizing public opinion. In algorithm design, platforms should balance commercial interests with social responsibility and explore ways to curb the spread of false information and the amplification of extreme speech.

The conflict between technology and regulation ultimately hinges on human support. When AI-generated content becomes difficult to thoroughly screen through technical means in the future, the public's critical digital literacy will emerge as a key factor in social defense. Social governance must include long-term educational investment to cultivate the public, especially young people, in the AI era's ability to discern information, understand algorithmic logic, and resist emotional manipulation. AI water armies are not only a technical issue, but also an economic and social one, requiring a systematic response across technology, law, platforms, and education to maintain a healthy online public opinion ecosystem.

DOI: 10.54254/2753-7048/2025.LD29977

## 5. Conclusion

This article focuses on the question of "Can AI trolls drive public opinion polarization?" Through a case study and mechanistic deduction of the 2024 AI troll army case in Sihong, Jiangsu, this study concludes that AI trolls are a new kind of force that manipulates public opinion, undergoing qualitative changes in terms of production costs, operational efficiency, and the deceptive nature of their content.

The Sihong case, as a prime example, while not directly employing generative AI, clearly reveals the industrialized, scaled infrastructure of trolls—an automated matrix comprised of thousands of cloud phones. This infrastructure, combined with the low-cost content production capabilities of generative AI, significantly lowers the economic barrier to public opinion manipulation. Deductions reveal that AI trolls exacerbate public opinion polarization primarily through two mechanisms: First, through a low-cost, large-scale "flood" of content, they create false public opinion, suppress moderate and dissenting views, and lead to apparent polarization; second, through interaction with platform recommendation algorithms, AI trolls can produce and disseminate large amounts of content that reinforces bias, thereby reinforcing filter bubbles and echo chambers and fostering attitudinal and emotional polarization. From a media economics perspective, AI water armies operate under the logic of the platform economy and the attention economy, exploiting extremely low information production costs to manipulate traffic and profit. This activity not only disrupts market order but also poses a serious challenge to social governance by polluting the information environment and exacerbating social tensions.

The conclusions of this study bear theoretical and practical implications. Theoretically, it suggests that AI technology, propelled by specific economic incentives and technological constructs, can act as a catalyst for public opinion polarization. On a practical level, this study advocates for an all-encompassing governance framework that transcends mere content moderation. First, at the legal level, as in the Sihong case, economic offenses such as illegal business operations should be used as a pretext to crack down on the black and gray industry chain of water armies and dismantle their infrastructure. Second, it is imperative to reinforce the primary accountability of platforms and advance reforms for algorithmic transparency in order to curb their propensity to amplify extreme rhetoric. Third, critical digital literacy education must be incorporated into long-term social governance strategies to enhance public media literacy.

However, this study is not without its limitations. Given the hidden nature of AI water armies, this paper primarily draws on public case studies and lacks empirical data on their actual communication effects. Future research should explore computational social science methods, such as simulation experiments or big data analysis, to more accurately measure the actual impact of AI-powered online campaigns on user attitudes and online polarization. Nevertheless, the risks of using AI technology to manipulate public opinion are already clear, and require continued attention and response from both academia and industry.

## References

[1] Shao, Z., Fang, Y., Zhang, Y., & Ciampaglia, G. L. (2024). Guarding the guardians: The challenge of detecting AI-generated disinformation. WIREs Data Mining and Knowledge Discovery, 14(3), e1537. https://doi.org/10.1002/widm.1537

[2] Prior, M. (2013). Media and political polarization. Annual Review of Political Science, 16(1), 101–127. https://doi.org/10.1146/annurev-polisci-052212-135306

[3] Pariser, E. (2011). The filter bubble: What the Internet is hiding from you. Penguin Press.

[4] Sunstein, C. R. (2017). #Republic: Divided democracy in the age of social media. Princeton University Press.

DOI: 10.54254/2753-7048/2025.LD29977

[5] Zhang, Y., Cheng, Y., & Yang, F. (2020). Astroturfing in online discussions: An integrative review of its definition, detection, and impact. First Monday, 25(2). https: //doi.org/10.5210/fm.v25i2.10057
[6] Acemoglu, D. (2024). The economics of generative AI. Communications of the ACM, 67(5), 44–48. https: //doi.org/10.1145/3644102
[7] DiMaggio, P., Evans, J., & Bryson, B. (1996). Have Americans' social attitudes become more polarized? American Journal of Sociology, 102(3), 690–755. https: //doi.org/10.1086/230995
[8] Vallone, R. P., Ross, L., & Lepper, M. R. (1985). The hostile media phenomenon: Biased perception and perceptions of media bias in coverage of the Beirut massacre. Journal of Personality and Social Psychology, 49(3), 577–585. https: //doi.org/10.1037/0022-3514.49.3.577
[9] Acemoglu, D., Ozdaglar, A., & Sastry, A. (2024). AI and social media: A political economy perspective. (NBER Working Paper No. 32549). National Bureau of Economic Research. https: //doi.org/10.3386/w32549
[10] Tucker, J. A., Guess, A., Barberá, P., Vaccari, C., Siegel, A., Sanovich, S., Stukal, D., & Nyhan, B. (2018). Social media, political polarization, and political disinformation: A review of the scientific literature. Hewlett Foundation.
[11] Bai, J., Li, Z., & Chen, H. (2024). How AI sources can increase openness to opposing views. Scientific Reports, 14(1), 10839. https: //doi.org/10.1038/s41598-024-61565-x