Scientific and Social Research, 2025, Volume 7, Issue 10
https://ojs.bbwpublisher.com/index.php/SSR
Online ISSN: 2981-9946
Print ISSN: 2661-4332

On Leveraging Multi-Agent Models in Countermeasures for Deepfake Detection and Mitigation: A Comparative Analysis of Social Media Platform Strategies

Yuanyuan Liu[1], Yu Zhang[1], Jicheng Sun[2]*

[1]School of Foreign Studies, China University of Petroleum (East China), Qingdao 266580, China
[2]School of Foreign Studies, Shandong University of Technology, Zibo 255000, China

**Corresponding author: Jicheng Sun, jichengsun@sdut.edu.cn*

Abstract: The proliferation of deepfake content on social media in recent years has posed significant threats to both individual and societal security. Consequently, devising effective countermeasures to curb the spread of deepfake information has become a critical challenge for social media platforms worldwide. This study aims to explore the propagation dynamics of deepfake information and evaluate the efficacy of various countermeasures by constructing a multi-agent model that integrates the SIR epidemiological model with the BA scale-free network theory. The research focuses on three distinct social media platforms in China—Douyin, Weibo, and Bilibili—as case studies. Through a series of simulation experiments, we compare the propagation patterns of deepfake content and analyze the performance differences of various countermeasures. The results indicate that, in terms of preventing the dissemination of deepfake information, the "preemptive defense" mechanism (exemplified by Douyin) proves to be the most effective in limiting the spread of deepfakes and ensuring timely counteractions. In contrast, the "post-verification" framework (exemplified by Weibo) is particularly effective in enhancing immunity against deepfake content. However, countermeasures based on self-media strategies that emphasize "emotion, viewpoints, and stances" (exemplified by Bilibili) demonstrate higher infection rates, weaker immunity, and longer response delays. The findings of this study offer valuable insights for developing more efficient and adaptive information governance strategies.

Keywords: Social media; Deepfake; Multi-agent; Simulation

Online publication: October 29, 2025

1. Background

In recent years, the rapid advancement of artificial intelligence (AI) technologies has led to an exponential growth in AI-generated content (AIGC) across various platforms, including social media and video-sharing websites.

AIGC encompasses multimodal content created by AI technologies, such as marketing materials, articles, product descriptions, images, audio, and videos [1].

According to incomplete statistics, the volume of deep synthetic content on internet platforms has increased exponentially, with a significant proportion comprising deepfake information [2]. Deepfake technology is based on "Generative Adversarial Networks" (GAN), a deep learning technique that enables the creation of hyper-realistic digital fakes in images, videos, and audio [3].

The creation of personalized and eye-catching misinformation can mislead public judgment, destabilize society, and contribute to the stigmatization problem on social media [4]. In Western countries, generative AI has been exploited maliciously to disseminate false political information, fabricate news, manipulate public opinion, and disrupt national elections, thereby negatively impacting political stability [5].

Current research on countering deepfake content on social media primarily focuses on social risks and legal regulations, identification and detection methods, and intervention strategies. **Figures 1** and **2** display 286 studies on the dissemination of deepfakes in social media, published between 2018 and June 2024 [6]. Among these, research on social risks and legal regulation is generally strategy-oriented, conducting qualitative studies on social risks, governance pathways, and multi-party collaboration [7–9]. In contrast, research on identification, detection, and intervention methods is more technology-oriented, focusing on optimizing detection models and improving dataset performance through quantitative research [10–11].

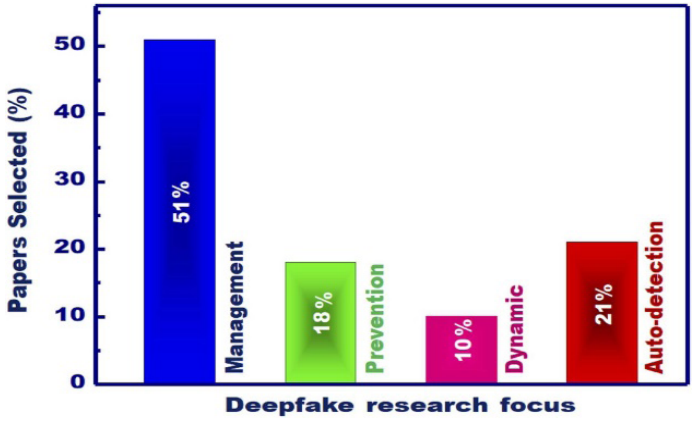

Figure 1. Selected papers with a focus on research [6]

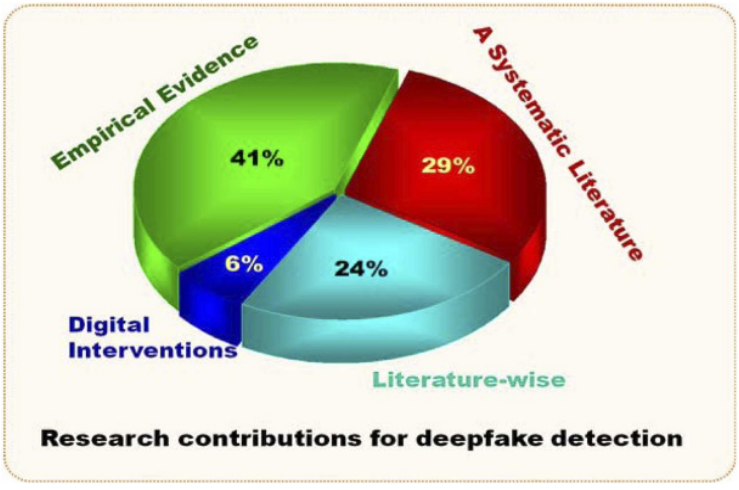

Figure 2. Research types of selected papers [6]

In conclusion, the current research on quantitative analysis of countermeasures against deepfakes on social media reveals significant gaps, thereby limiting the use of scientific and objective methods to assess the effectiveness of these countermeasures. To address this research gap, the study will conduct simulation experiments and effectiveness evaluations of both the propagation and countermeasures of identical deepfake content across different social media platforms. The ultimate goal is to optimize existing countermeasures and develop solutions to address the prevailing challenges in this domain.

2. Model construction

Social media deepfake countermeasures require hierarchical governance, comprising three components: system behavior, structure, and environment [3]. The system structure represents a relatively stable, goal-aligned element organization, while system behavior encompasses agent interactions (including needs, motives, stimuli, goals, and feedback) that achieve functional objectives. The system environment includes all external interacting factors [12].

This study employs Agent-Based Modeling (ABM) to simulate emergent user behaviors, as ABM effectively reveals macroscopic effects from individual behaviors in complex adaptive systems [13]. Individual users are abstracted as agents to analyze the effectiveness of the platform against deepfake propagation.

Network Structure: Social media networks exhibit BA scale-free properties, characterized by a power-law degree distribution—most nodes have low degrees, while a few have high degrees [14]. This study uses BA scale-free networks to simulate real social media environments [15].

Propagation Behavior: Deepfake information spreads "virally" [16]. The SIR epidemiological model categorizes nodes as S (susceptible), I (infected), or R (immune/recovered), effectively modeling information diffusion on networks [9].

Considering user diversity, interaction heterogeneity, and relationship complexity, the modeling approach follows "Agent-Based Modeling of Weibo Group Event Propagation" (**Figure 3**) [17].

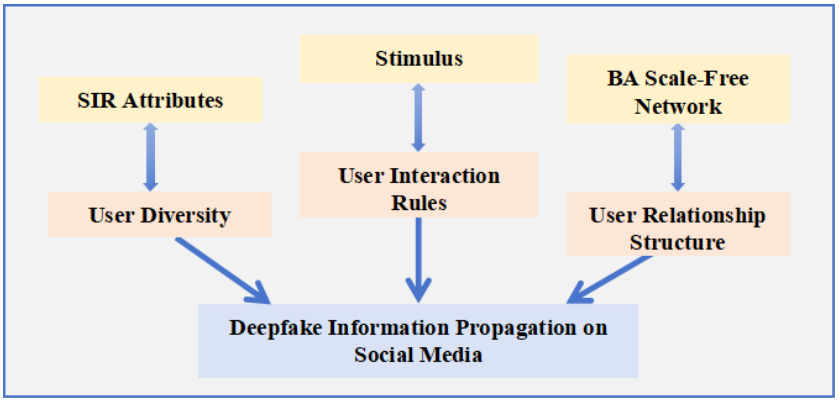

Figure 3. Deepfake information propagation model on social media

3. Simulation experiment

This study employs an ABM model integrating the SIR model with BA scale-free networks to evaluate deepfake countermeasures across social media platforms, using the "Xiamen Representative Bai Jie" deepfake case.

Three platforms with distinct strategies were examined:

Douyin: Preemptive defense with AI-generated content alerts, posting restrictions, and virtual human live-streaming controls.

Weibo: Post-verification via official rumor refutation accounts and third-party co-governance platforms.

Bilibili: Emotion-driven countermeasures leveraging anti-fraud creators and rapid recommendation algorithms to question and expose deepfakes.

Data were collected from all three platforms and simulated using NetLogo 6.3.0.

3.1. Data collection

Data collection employed a "dual snowball" sampling method [17]. First, event-related keywords were used to retrieve all relevant posts within the specified timeframe across platforms. Second, activities such as reposting, quoting, and replying were tracked to capture missed keyword-related content. Third, all posts from identified users during the event period were collected to ensure comprehensive coverage.

Given the limited deepfake videos (mean = 3), K-shell decomposition for influence calculation was unnecessary. Table metrics (followers, favorites, likes, reposts, and comments) represent aggregated values for fake and debunking videos across platforms. View counts for Douyin and Weibo were estimated via sampling due to data unavailability.

Table 1. Propagation data of deepfake videos involving "Xiamen Representative Bai Jie"

Number of fake videos	Platform	Favorites	Likes	Reposts	Comments	Views
2	Douyin	2	48	5	1	8,000
2	Weibo	0	1512	364	2227	50,000
1	Bilibili	15	196	81	Banned	18,000

Table 2. Propagation data of debunking videos involving "Xiamen Representative Bai Jie"

Number of debunking videos	Platform	Favorites	Likes	Reposts	Comments	Views
7	Douyin	7	5280	4	44	40,000
4	Weibo	0	6441	128	420	100,000
2	Bilibili	205	6793	128	378	56,000

Note: Likes = likes + 10× coins; Weibo favorites = 0 (feature unavailable); Total fake content impressions ≈ 1.16 million

3.2. Parameter settings

Parameters for initial outbreak scale, virus check frequency, and transmission/recovery/resistance opportunities were derived from platform data.

Notably, the virus check frequency represents the intervals at which infected nodes are monitored and managed within the simulation framework. In the context of computer viruses and information dissemination models, fixed inspection frequencies are often employed to simulate the cyclical scanning or monitoring actions undertaken by defense mechanisms. This experiment was fixed at 10 to ensure temporal consistency and simulation stability across platforms.

The parameter delineating transmission opportunity signifies the propensity for deepfake information to spread within a social network. The Propagation Opportunity (PO) is computed as follows:

$$PO = 10 * \frac{Likes+Favorites+Comments+Shares}{Impressions} \quad (1)$$

User interactions (likes, favorites, comments, shares) indicate propagation potential [9]—normalization by impressions controls for user base variance. The amplification factor (×10) reflects empirical evidence that false information spreads more rapidly than debunking content [17].

The Recovery Opportunity (RO) refers to the likelihood that users will recover from an infected state after being exposed to debunking information. RO is calculated as follows:

$$RO = \frac{Likes+Favorites+Comments+Shares}{Impressions} \quad (2)$$

In formula (2), the debunking video interactions measure penetration and user acceptance [18]. Similarly, equation (2) normalizes the interaction volume by impressions to ensure consistency and equitable evaluative criteria. It is noteworthy to acknowledge that, in contrast to false information, debunking content often encounters impediments characterized by "delayed initiation and diminished attention." No amplification is applied, reflecting lower engagement with corrective content [19].

The BA scale-free network with 100 initial nodes balances scale-free properties with computational efficiency. Resistance Opportunity (ReO) is given by:

$$ReO = \frac{Number\ of\ Debunking\ Videos * Impressions}{100 * Number\ of\ Fake\ Videos * Impressions} \quad (3)$$

The video ratio reflects the quantitative advantage of debunking information [20]. Impressions are adjusted for exposure impact; division by 100 scales the results appropriately.

3.3. Model execution

Substitute the data from **Table 1–2** into Formula 1–3 to determine the initial value. The results are shown in **Table 3**.

Table 3. Initial values for the ABM deepfake information propagation model

Virus check frequency	PO (%)	RO (%)	ReO (%)	Initial degree distribution
10	7	13.3	17.5	100
10	8.2	7	40	100
10	9.3	13.4	6.2	100

Based on the initial values provided in Table 3, the data were then input into the deepfake information propagation model, and simulation experiments were conducted using NetLogo. The experimental results are detailed below:

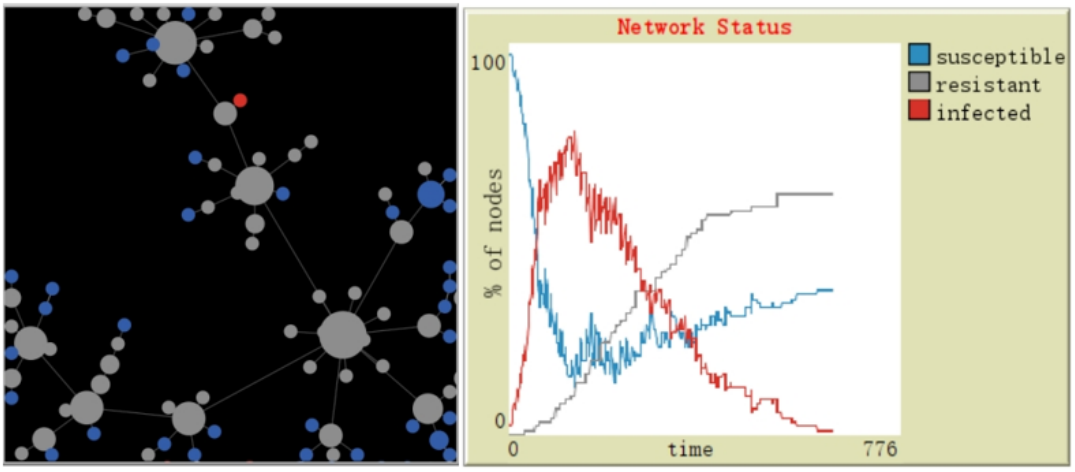

Figure 4. Simulation of deepfake information propagation on Douyin

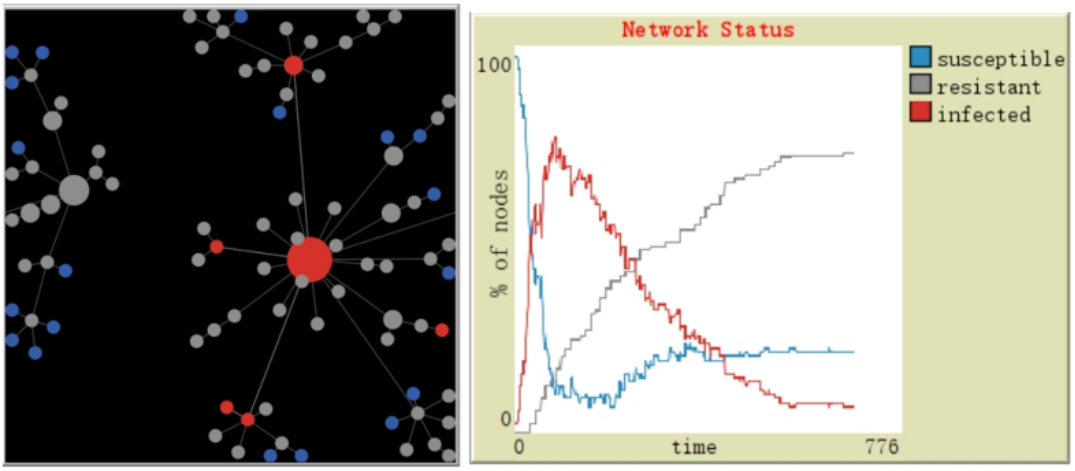

Figure 5. Simulation of deepfake information propagation on Weibo

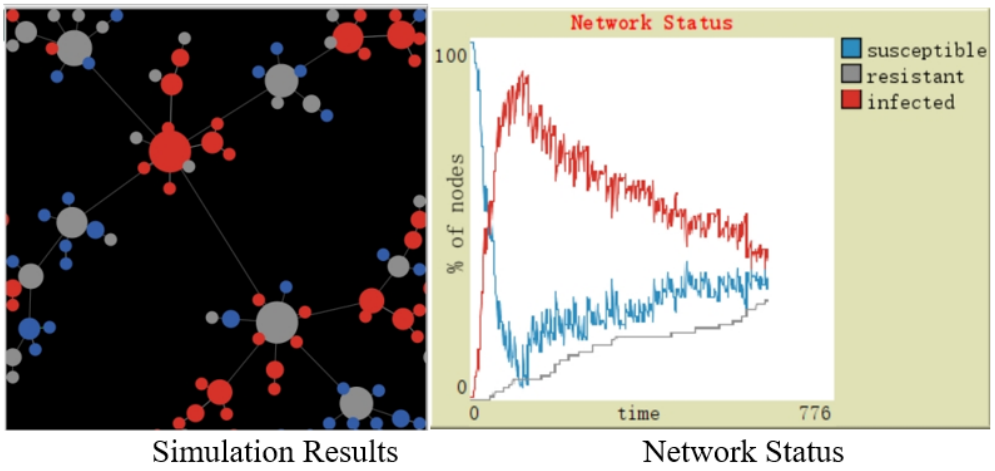

Figure 6. Simulation of deepfake information propagation on Bilibili

4. Results and Discussion

4.1. Results

The simulation experiment results indicate that the structure of various countermeasure systems significantly affects their response speed and effectiveness in mitigating the spread of deepfake information. The specific analysis is as follows:

(1) Number of Infections and Propagation Control: By comparing the peaks of the red curves, it is observed that the number of infections is lowest on the Douyin platform, whereas the peak is highest on Bilibili. Furthermore, from the endpoints of the red curves, it is evident that Douyin was the fastest to reduce the infection value to the minimum, achieving timely control of deepfake propagation. In contrast, Weibo demonstrates moderate control effectiveness, while Bilibili exhibits the slowest control effects.

(2) Trends in Immune Population: The trend of the gray curve indicates that Weibo and Douyin exhibit strong immunity. However, Bilibili, which employs an "emotional stance" countermeasure strategy, exhibits lower immunity and slower growth despite having a smaller susceptible population, resulting in a faster short-term spread of deepfake information. Overall, its effectiveness is inferior to the other mechanisms.

(3) Analysis of Susceptible Population Characteristics: By comparing the lowest points of the blue curves representing the susceptible population, it is found that Weibo's countermeasure strategy results in the shortest feedback time, quickly reducing the susceptible population to a minimum. Douyin and Bilibili follow behind. However, in subsequent stages, the susceptible population on all three platforms demonstrates slight increases before eventually leveling off.

(4) Final Node Distribution: When deepfake propagation reached a stable state across the three platforms, the study counted the number of agents with different attributes, specifically the SIR nodes in the simulation experiment (**Figure 7**). The results show that Douyin has the fewest susceptible individuals and the most immune individuals; Bilibili has the most susceptible individuals and the fewest immune individuals; Weibo has the highest number of immune individuals.

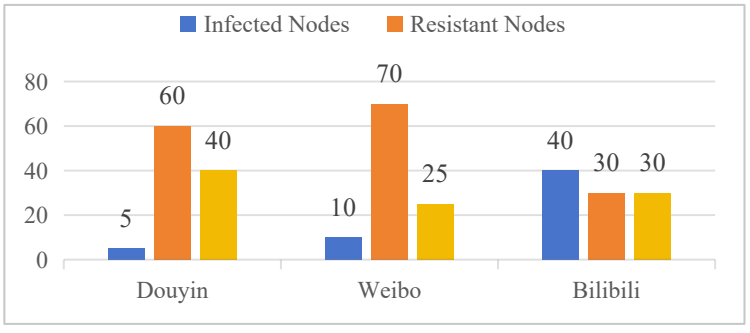

Figure 7. Final node distribution across platforms (100% Normalized)

The peak infection rate, time proportion, final infected nodes, final resistant nodes, and final susceptible nodes for deepfake information propagation are summarized as follows (**Table 4**):

(1) The Douyin platform demonstrates the least extensive dissemination of deepfake content, coupled with the most expedited recovery process, with a short feedback delay time.

(2) The Weibo platform was the first to achieve immunity to deepfake information, with the highest proportion of user base that eventually develops resistance against such content.

(3) The Bilibili platform exhibits a notably elevated infection rate and the weakest efficacy in developing

immunity, accompanied by a longer feedback delay time.

Table 4. Peak infection rates and final node statistics

Platform	Peak infection rate (%)	Time proportion (%)	Final infected nodes (%)	Final resistant nodes (%)	Final susceptible nodes (%)
Douyin	70	20	5	60	40
Weibo	70	15	10	70	25
Bilibili	90	10	40	30	30

4.2. Discussion

Effectiveness of the "Preemptive Defense" Strategy: The Douyin platform, using a "preemptive defense" countermeasure structure, demonstrated significant advantages in terms of the number of infections, recovery speed, and feedback delay time. It indicates that early technical alerts and warning systems can effectively mitigate the further spread of misinformation. The "preemptive defense" structure outperforms alternative countermeasure mechanisms in terms of both efficiency and efficacy.

Immunity of the "Post-Verification" Strategy: The Weibo platform's adoption of a "post-verification" mechanism, incorporating third-party oversight and fact-checking procedures, facilitated a rapid emergence of immunity and a rapid enhancement of user immunity over a relatively short time. Although the initial immune response was slightly slower compared to Douyin, Weibo eventually achieved the highest level of user immunity, highlighting the critical role of verification and correction mechanisms in managing deepfake information.

Limitations of the "Emotional Stance" Strategy: In contrast, Bilibili's countermeasure strategy, which relied on "emotional stance", was the least effective. Despite having a smaller susceptible population, Bilibili exhibited a higher infection rate, lower immunity effectiveness, and prolonged feedback delay times. These results indicate that relying solely on self-media recommendation algorithms and emotional interactions is insufficient to curtail the spread of deepfake content and may lead to delayed immunity. Instead, a multifaceted approach is essential for developing comprehensive strategies to address the challenges posed by the dissemination of deepfake information.

Dynamic Interactivity of Media Countermeasures: Overall, the countermeasures employed by social media platforms exhibited strong dynamism and openness. The varying effectiveness of different combinations of countermeasures across platforms underscores the importance of considering platform structure, user behavior, and external regulatory environments in their design. Therefore, when devising countermeasures for social media platforms, a multi-layered, multi-strategy, comprehensive approach should be adopted.

In summary, media countermeasures operate within an open and dynamic energy field, with their effectiveness contingent upon the interaction between system structure, behavioral patterns, and the external environment. This study innovatively applies the agent-based modeling (ABM) approach to the study of deepfake countermeasures, revealing that the structure of countermeasure systems significantly influences the response speed and efficacy of managing deepfake content. Quantitative analysis indicates that the "preemptive defense" structure achieves the fastest suppression of spread, while the "post-verification" structure demonstrates robust immunity.

Funding

This study was supported by the Humanities and Social Sciences Project Grants by the Ministry of Education of China (24YJAZH094) and 2025 General Program of the National Social Science Fund of China (25BXW099).

Disclosure statement

The authors declare no conflict of interest.

References

[1] Wu J, Gan W, Chen Z, et al., 2023, AI-Generated Content (AIGC): A Survey. arXiv, accessed on October 2, 2025. https://arxiv.org/abs/2304.06632

[2] Tsinghua University Institute for Artificial Intelligence, 2022, Top 10 Trends in Deep Synthesis. Retrieved on October 2, 2025, https://www.digitalelite.cn/h-nd-3039.html

[3] Cai S, 2020, The Technical Logic and Legal Changes of Deepfake. Political and Legal Review, 2020(3): 131–140.

[4] Dash B, Sharma P, 2023, Are ChatGPT and Deepfake Algorithms Endangering the Cybersecurity Industry? A Review. International Journal of Engineering and Applied Sciences, 10(1): 4–5.

[5] Chen Y, Wang W, 2023, From "Immediate Implementation" to "Gradual Improvement": Constructing the Governance System of Generative AI. E-Government, 2023: 15–27.

[6] El-Sayed Atlam M, Malik Almaliki G, Ghada Elmarhomy, et al., 2025, SLM-DFS: A Systematic Literature Map of Deepfake Spread on Social Media. Alexandria Engineering Journal, 2025(111): 446–455.

[7] Wang G, Zhang Z, 2024, Characteristics, Risks, and Countermeasures of Deepfake Technology. New Media and Network, 1(2): 40–51.

[8] Zhang X, Wang R, Ma Y, 2024, Challenges, Opportunities, and Strategies in Fake Information Governance Under the AIGC Context. Information Science, 1–23, accessed on October 2, 2025, https://link.cnki.net/urlid/22.1264.G2.20241111.1002.024

[9] Zhao L, Wang J, Cheng J, et al., 2012, Rumor Spreading Model with Consideration of Forgetting and Remembering Mechanisms in Social Networks. Physica A: Statistical Mechanics and its Applications, 391(7): 2444–2453.

[10] Sadiq S, Aljrees T, Ullah S, 2023, Deepfake Detection on Social Media: Leveraging Deep Learning and FastText Embeddings for Identifying Machine-Generated Tweets. IEEE Access, 2023(11): 95008–95021.

[11] Khan AA, Chen YL, Hajjej F, et al., 2024, Digital Forensics for the Socio-Cyber World (DF-SCW): A Novel Framework for Deepfake Multimedia Investigation on Social Media Platforms. Egyptian Informatics Journal, 27(1): 1–12.

[12] Mao H, 2013, Core Elements of the Cross-Strait Dialogue System, thesis, Fujian Normal University.

[13] Jiang S, Han Z, 2011, Agent-Based Modeling Methods in Complex Systems Research. Journal of the University of Shanghai for Science and Technology, 33(2): 124–129.

[14] Pan J, Shen H, Chen Z, 2018, A Model of Group Event Microblog Propagation Based on Agent and K-Core Decomposition. Information Science, 36(2): 125–131.

[15] Sun R, Luo W, 2014, Rumor Spreading Model in Scale-Free Networks with Non-Uniform Propagation Rates. Complex Systems and Complexity Science, 11(3): 6–11.

[16] Liu C, Chen M, 2023, User Perception and Interaction Behaviors of False Information Based on Cue Utility Theory: Focused on Deepfake Information. Journal of Information Science, 42(10): 96–104.

[17] Vosoughi S, Roy D, Aral S, 2018, The Spread of True and False News Online. Science, 359(6380): 1146–1151.

[18] Zhao L, Wang J, Chen Y, et al., 2012, SIHR Rumor Spreading Model in Social Networks. Physica A: Statistical Mechanics and its Applications, 391(7): 2444–2453.

[19] Friggeri A, Adamic L, Eckles D, et al., 2014, Rumor Cascades. Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media, 101–110.

[20] Granell C, Gomez S, Arenas A, 2013, Dynamical Interplay Between Awareness and Epidemic Spreading in Multiplex Networks. Physical Review Letters, 111(12): 128701.