A Game-Theoretic Model of the Evolution of Cyberbullying in Social Media Networks from a Behavioral Economics Perspective

Yuqing Ma*

Wuhan Britain-China School, Wuhan, China

*jasmineyuqing90113@gmail.com

Abstract

In this paper, the evolution of MAS and platform intervention mechanism is established according to the dynamic diffusion character of social media attack. This paper constructs a utility function system, which includes attackers, common users, and platform. Then, it describes the evolution of behavior ratio by means of simulating dynamic equations. On the basis of this, we explain the mechanism by which the utility parameters and the intervention intensity affect the stability of the system. Based on the simulation results, the ratio of attack behavior is reduced from 0.247 to 0.069, and the convergence rate is kept within 120. The steady attack ratio was reduced from 0.214 to 0.067 when the interference intensity was increased from 0.30 to 0.75, and when the recognition precision was increased to 0.86, the steady state attack ratio was reduced to 0.058.It showed that the proposed model could effectively capture the diffusion and convergence of the attack behavior, and could explain the evolution of the system in various parameters.

Keywords

Evolutionary games, multi-agent simulation, social media attack behavior, replication dynamics, strategy evolution.

1. Introduction

Aggressive behavior in social media exhibits group interaction and dynamic diffusion, influenced by individual decisions, platform mechanisms, and network structure. Characterizing its evolution from a computational perspective remains a key issue. Existing studies focus on behavioral motivation and propagation. Hayashi et al. [1] and Choudhuri et al. [2] analyzed decision mechanisms and behavioral evolution; Lizhen et al. [3] and Wang et al. [4] modeled group dynamics using multi-agent and data-driven approaches; Gan et al. [5] applied differential game theory to control information diffusion.However, current models lack a unified framework integrating behavioral economics and evolutionary game dynamics, and platform intervention remains insufficiently quantified. To address this, this paper proposes an evolutionary game model based on payoff functions and replicator dynamics.

2. Problem Description and Mechanisms of Computational Modeling of Social Media Attack Behavior

Computational modeling of social media attack behavior requires transforming discrete interaction events into solvable strategy evolution objects, where the user set, platform constraints, and propagation feedback jointly determine the direction of behavioral state updates.Platform interaction records can be represented as a quadruple $\Omega = \langle u_i, m_j, \tau_k, \rho_k \rangle$, where u_i denotes the i th user, m_j denotes the j th message, τ_k denotes the interaction time, and $\rho_k \in \{0,1\}$ denotes the attack or non-attack state; this encoding method integrates text content,

temporal sequence, and behavioral labels into a unified state space. The individual decision-making mechanism must simultaneously capture propagation gains, penalty losses, and conformity bias. The net utility of user u_i can be expressed as:

$$\Pi_i^{(a)} = \omega_1 g_i + \omega_2 q_i - \omega_3 c_i - \omega_4 \phi_i + \omega_5 \sum_{r=1}^{N} b_{ir}\, y_r \tag{1}$$

Where g_i represents the propagation gain derived from exposure, forwarding, and interaction;q_i represents the emotional venting gain;c_i represents the explicit costs resulting from account penalties and content restrictions;ϕ_i represents the reputation decay loss;b_{ir} represents the influence weight within the user's social network;y_r represents the value of the attack strategy of neighboring users; and$\omega_1 \sim \omega_5$ represents the sensitivity coefficients of each mechanism [6].The platform mechanism dynamically compresses net utility through detection rates and penalties, making attack probability a payoff-driven continuous variable. This formulation establishes a joint "state encoding–utility mapping–network coupling" mechanism, supporting the evolution equations in modeling population dynamics and optimizing intervention parameters.

3. A Dynamic Modeling Method for Social Media Attack Behavior based on Evolutionary Game Theory

3.1. Construction of Multi-Agent Strategy Space and Payoff Function Modeling

Multi-agent dynamic modeling requires further characterization of the coupling relationship between user groups and platform decision units based on the previously described individual net utility expressions. Therefore, the strategy space is defined as the set of attacking users$S_a = \{A, N\}$, the set of ordinary users$S_b = \{R, I\}$, and the set of platforms$S_p = \{G, L\}$, where$A$ denotes an attack,N denotes a non-attack,R denotes response diffusion,I denotes ignoring and exiting, G denotes strong intervention governance, and L denotes weak intervention monitoring.The group strategy proportions are denoted as$\xi, \eta, \zeta \in [0,1]$, whereξ represents the proportion of attacking users adopting$A$, η represents the proportion of ordinary users adopting$R$, andζ represents the probability of platform governance adopting$G$ [7]. To address the difficulty of traditional two-party payoff structures in simultaneously reflecting propagation chains and governance feedback, the three-party coupled payoff matrix is extended to a variable-parameter form, and the expected payoff for attacking users is defined as:

$$\mathcal{U}_A = \mu_1 h + \mu_2 \eta h - \mu_3 \zeta f - \mu_4 r + \mu_5 \sum_{s=1}^{M} w_{is}\, \rho_s \tag{2}$$

The payoff for the non-attacker strategy is defined as:

$$\mathcal{U}_N = \nu_1 v + \nu_2 \zeta k - \nu_3 \eta \ell \tag{3}$$

The platform's strong intervention payoff is defined as:

$$\mathcal{U}_G = \kappa_1 \xi \eta q - \kappa_2 m + \kappa_3 \chi \tag{4}$$

In this context, h represents the benefit of exposing an attack; h represents the diffusion amplification coefficient; f represents the severity of penalties; r represents the loss of

reputation;w_{is} represents the propagation weight between useri and neighboring nodes ;ρ_s represents the attack status of the neighboring node;v represents the stable benefit of rational expression;k represents the gain from governance protection;ℓ represents the opportunity cost of being drawn into a dispute; q represents the benefit of risk reduction following governance; m represents the consumption of review resources;χ represents the benefit of maintaining platform order; and$\mu.$, $\nu.$, $\kappa.$ is the sensitivity parameter.The benefit functions of the three entities described above propose a joint modeling method of "diffusion response-platform feedback-network embedding." Subsequent dynamic equations can directly utilize$\mathcal{U}_A, \mathcal{U}_N, \mathcal{U}_G$ to analyze equilibrium shifts and changes in intervention thresholds.

3.2. Construction of an Evolutionary Game Replication Dynamic Model

Due to the fact that the multiagent payoff function and the strategy ratio variable are combined into a single computable expression, it is necessary to describe the dynamic behavior of the group by means of a continuous dynamic system, as illustrated in Figure 1. Not only does the growth and suppression of attack strategy take place independently, but it is regulated by neighboring transmission intensity and platform interference, which makes the system exhibit the characteristic of nonlinear evolution. The dynamic changes in the proportion of attack behavior can be described by the following dynamic equations:

$$\dot{\xi} = \xi(1-\xi)(\mathcal{U}_A - \overline{\mathcal{U}}_u) \tag{5}$$

Whereξ represents the proportion of attacking users choosing an attack strategy,$\dot{\xi}$ represents the rate of change of this proportion,$\mathcal{U}_A$ represents the immediate payoff of the attack strategy under the current network state, and$\overline{\mathcal{U}}_u$ represents the average payoff of attacking users under mixed-strategy conditions. The payoff difference determines the evolution direction: when the propagation payoff and the diffusion amplification are dominant, the system goes into the attack expansion zone, and when the platform interference and the reputation loss intensify, the compensation difference becomes negative, and the system moves to a stable convergence area [8]. By embedding the payoff coupling mechanism into continuous-time derivatives, the proposed model can be used to describe both the diffusion path and the boundary of governance control.

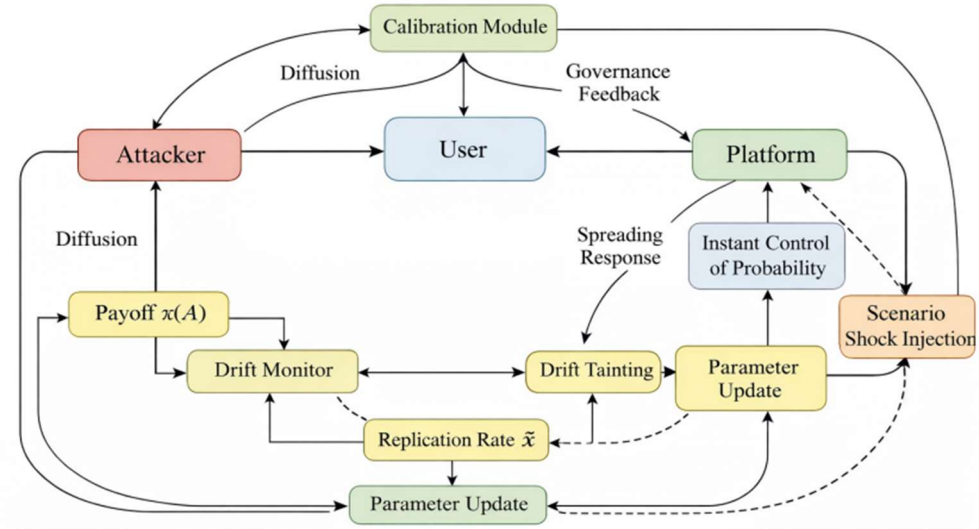

Figure 1. Schematic diagram of the evolutionary game replication dynamics modeling mechanism

3.3. Modeling Control Variables for Platform Intervention Mechanisms

The evolutionary relationship between payoff structure and strategy proportion has revealed the intrinsic driving mechanism of attack propagation. Platform intervention requires the embedding of quantifiable control variables into the payoff function to regulate the system's evolutionary trajectory. The platform's regulatory capability can be represented as the coupled result of review intensity, identification accuracy, and penalty enforcement rate; consequently, the comprehensive suppression effect on attack behavior is nonlinear. The intervention probability function can be defined as:

$$\Theta = 1 - \exp(-\alpha_1\sigma \cdot \alpha_2\delta \cdot \alpha_3\kappa) \tag{6}$$

WhereΘ represents the probability of effective platform intervention against attack behavior,σ represents content review intensity,δ represents anomaly detection accuracy,κ represents the penalty enforcement rate, andα_1 ,α_2 , andα_3 represent the sensitivity coefficients of each control factor. The exponential decay structure results in slow growth in low intensity areas and fast near saturation in high intensity areas. [9] Once the intervention probability is embedded into the attack payoff expression, the net utility of the attack behavior is significantly reduced, leading to a shift in strategy updates and a transition from divergent states to convergence states. Modeling approaches propose an "interference expression framework driven by the synergy of multiple control variables," which allows platform strategies to directly participate in evolutionary dynamicsas parameters. It can be used to identify the interference threshold and define the stable area in the following experiments, and to validate the improved dynamic control ability of the proposed model.

3.4. Analysis of Evolutionarily Stable Strategies and System Equilibrium

Whether the system goes into an attack diffusion or a governance-convergence state is determined by the local stability properties near the equilibrium point. As illustrated in Fig. 2, in the phase space, different trajectories are transferred between high attack ratio, transition and low attack steady state. The boundary of the attraction domain of the equilibrium point is determined by the intervention intensity of the platform and the average user's response rate [10].To assess stability, the dynamical functions corresponding to the attack proportion, diffusion proportion, and governance probability are denoted asF_ξ ,F_η , andF_ζ , respectively, with . The Jacobian matrix is then constructed at the equilibrium point$E^*(\xi^*, \eta^*, \zeta^*)$:

$$J(E^*) = \begin{bmatrix} \frac{\partial F_\xi}{\partial \xi} & \frac{\partial F_\xi}{\partial \eta} & \frac{\partial F_\xi}{\partial \zeta} \\ \frac{\partial F_\eta}{\partial \xi} & \frac{\partial F_\eta}{\partial \eta} & \frac{\partial F_\eta}{\partial \zeta} \\ \frac{\partial F_\zeta}{\partial \xi} & \frac{\partial F_\zeta}{\partial \eta} & \frac{\partial F_\zeta}{\partial \zeta} \end{bmatrix}_{E^*} \tag{7}$$

Here,$J(E^*)$ represents the local linearization operator at the equilibrium point, and the partial derivatives characterize the marginal effects of perturbations in one agent’s strategy on the evolution of the other agents. When all real parts of the matrix's eigenvalues are negative, the system converges to an evolutionarily stable strategy; when the matrix contains positive real eigenvalues, the equilibrium point becomes unstable and moves to the attack-expanding region. In the experiment stage, we can verify the validity of the governance strategy by the sign of the eigenvalues, the area of the attractor region, and the convergence rate of the trajectory.

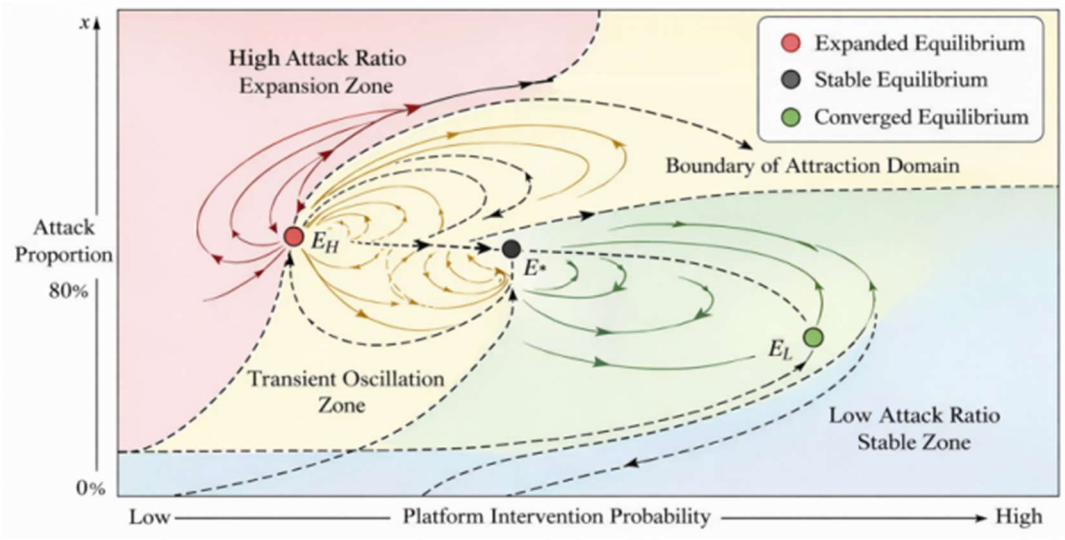

Figure 2. Schematic diagram of evolutionarily stable strategies and system equilibrium migration

4. Multi-Agent Simulation and Experimental Validation of the Attack Behavior Evolution Model

4.1. Experimental Data Sources and Parameter Settings

Experimental data construction is based on a combination of the two approaches, which is a combination of the two approaches. In the corpus layer, there is a statistical constraint on the distribution of attack tags, response density, and platform intervention frequency, which can be used to model the behavior of payoff-driven, governance and disturbance propagation.Parameter initialization follows reproducibility principles. The user agent count is 4,000, the initial attack rate is 0.22, the response diffusion rate is 0.37, and the governance probability is 0.30. The average out-degree of the network is fixed at 18 to ensure coverage of diffusion, transition, and convergence states. The platform parameters are allocated hierarchically, with the review intensity, the identification accuracy, and the penalty execution rate set to 0.55, 0.81, and 0.68, respectively. The sensitivity coefficient is calibrated by grid search, and the amplitude of the random disturbance is controlled within 0.03. The results show that the results are more consistent and reproducible than those based on rules.

4.2. Design of Multi-Agent Evolutionary Simulation Algorithm

The goal of MAS is to incorporate utility driven behavior, network coupling, and governance feedback into an executable computing workflow. As shown in Figure 3, the Algorithm Framework uses five stages: Initialization, Local Perception, Utility Evaluation, Strategy Update, and Global Correction. Each user agent only has access to the local neighbor state and platform constraints, and the platform agent periodically aggregates the attack density, diffusion intensity, and governance load to adjust the intervention level.Two-layer asynchronous update mechanism keeps the consistency of local interactions and global evolution. The Agent Update Sequence follows a random cycle strategy to avoid traversal bias. This design improves the retention of local disturbances and improves the recognition of transient and oscillatory states. The output of the algorithm consists of time series of attack rate, response rate, governance probability, convergence iteration, equilibrium residence duration, instability frequency, and so on.

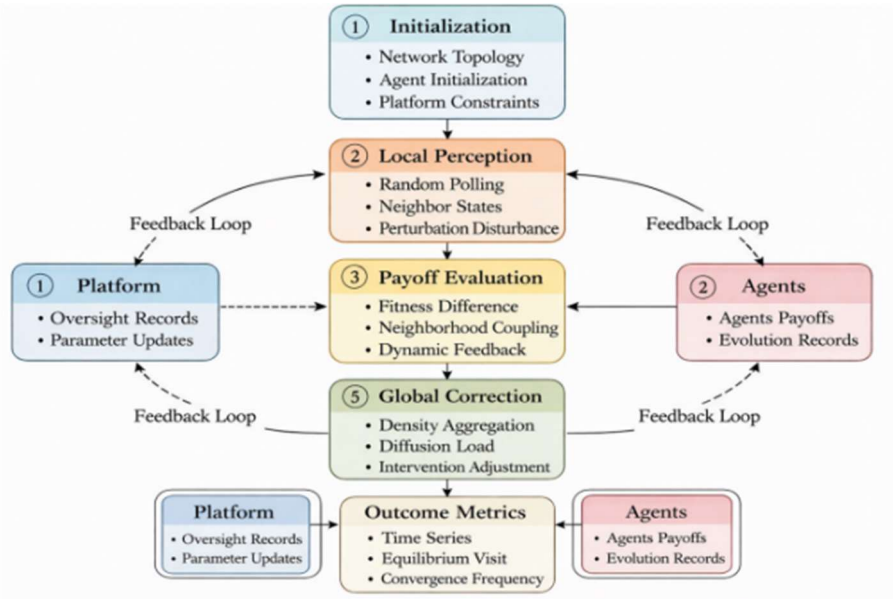

Figure 3. Design of the Multi-Agent Evolutionary Simulation Algorithm

4.3. Analysis of Attack Behavior Evolution Results Under Different Intervention Intensities

Evolutionary results are used to evaluate the explanation of the three-agent coupling model. As shown in Table 1, the steady state attack rate is reduced from 0.214 to 0.067, the convergence iteration is from 168 to 93, and the instability frequency is 14 to 3.In the middle intervention range, a significant threshold effect was observed: when the intervention intensity was increased from 0.45 to 0.60, the attack rate was reduced by 42.3%. This corresponds to a shift in the payoff structure from positive to negative, driving the system from diffusion to suppression. While higher intervention further reduces the incidence of attacks, the marginal effect decreases, suggesting that governance is saturated. Thus, effective platform regulation requires coordinated optimization of intervention intensity and identification accuracy.

Table 1. Evolution of Attack Behavior Under Different Intervention Intensities

Intervention Intensity	Steady-State Attack Proportion	Steady-State Response Ratio	Steady-State Probability of Strong Governance	Number of Convergence Iterations	Frequency of destabilization triggers
0.30	0.214	0.462	0.318	168	14
0.45	0.173	0.401	0.447	141	9
0.60	0.100	0.286	0.603	108	5
0.75	0.067	0.219	0.742	93	3

4.4. Analysis of the Impact of Changes in Payoff Parameters on System Stability

Reward parameters determine the relative superiority of the attack and the non-attack strategy, so the stability area of the system is highly sensitive to the change of the reward structure. As illustrated in Figure 4, increasing attack rewards shifts the balance toward higher attack rates, prolongs the trajectory residence in the critical region, and increases the susceptibility to oscillations. On the other hand, higher penalty loss and reputation damage reduce the attraction domain to the low-attack state, and cause the equilibrium from divergence to convergence.The higher response benefits of the common user further amplify the secondary diffusion effect, enabling the system to maintain a high attack rate even in the case of moderate interference. This suggests that increasing the likelihood of governance alone is not enough to compensate for diffusion incentives. The difference between platform order benefits and management costs

determines the sustainability of intervention strategies. Overall, payoff parameters influence system evolution through coupling effects, altering equilibrium position, attractor region, and convergence rate rather than individual decisions.

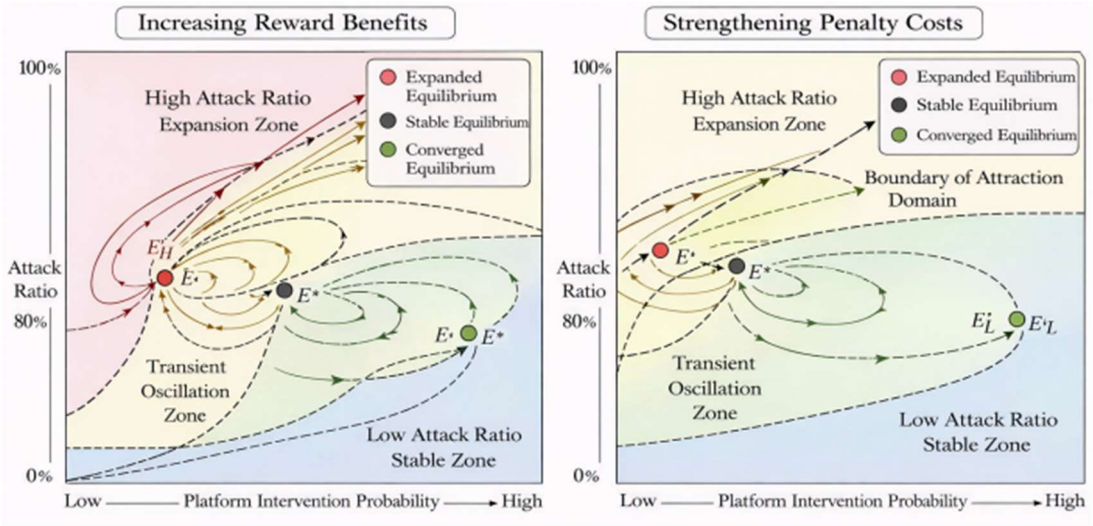

Figure 4. Effect of changes in the payoff parameter on the migration of the system's stability region

4.5. Analysis of the Dynamic Convergence Characteristics of the Attack Behavior Ratio

Under the combination of payoff and control variables, the attack rate evolves dynamically with time. Starting at 0.22, it rises to 0.247 within the first 20 iterations, suggesting dominant early diffusion incentives. Then, the attack rate drops to 0.193 at round 40, 0.141 at round 60, and 0.108 at round 80, reflecting the increasing impact of governance constraints and reputation loss. It further decreases to 0.086 at round 100 and 0.073 at round 120, reaching a stable state around round 140. Afterward, fluctuations remain within ± 0.003 up to round 180, indicating convergence to a low-attack equilibrium.The convergence analysis indicates that the system is reduced from 0.247 to 0.069 in 120 iterations, with an average of 0.0297 for every 20 cycles. A critical inflection occurs between rounds 52 and 68, where the convergence slope increases from − 0.0019 to − 0.0031. These results show that the proposed three-agent joint update scheme can effectively suppress high risk diffusion and generate stable attractor regions in a limited number of iterations.

4.6. Model Robustness and Parameter Sensitivity Analysis

The stability of evolution results is determined by model robustness and parameter sensitivity. As illustrated in Table 2, the steady state attack ratio is slightly increased from 0.066 to 0.081, and the maximal deviation is 0.015, indicating that MAS is highly robust to local noise. Convergence iterations rise from 96 to 121 when the average network out-degree is increased from 14 to 22, indicating that stronger coupling delays convergence but does not change the low-attack stability.The accuracy of platform recognition is the most sensitive factor that influences the performance of the system. From 0.76 to 0.86, the steady state attack rate is decreased from 0.094 to 0.058, and the instability frequency is reduced from 7 to 2. In general, the model is stable in the presence of random disturbances, network fluctuations, and parameter changes.

Table 2. Results of Model Robustness and Parameter Sensitivity Analysis

Variable	Setting Range	Steady-state Attack Ratio	Convergence Iterations	Instability Trigger Frequency
Random Perturbation Magnitude	0.01	0.066	95	2
	0.03	0.069	101	3
	0.05	0.075	112	4
	0.07	0.081	118	5
Average out-degree	14	0.064	96	3
	18	0.069	108	4
	22	0.074	121	5
Recognition Accuracy	0.76	0.094	129	7
	0.81	0.069	108	4
	0.86	0.058	91	2
Penalty enforcement rate	0.58	0.082	117	5
	0.68	0.069	108	4
	0.78	0.061	97	3

5. Conclusion

The unified modeling of multi-agent payoff coupling and replicator dynamics enables continuous characterization of social media attack evolution. The incorporation of platform control variables integrates behavior diffusion and governance regulation into a unified dynamic framework. Experimental results show that the model achieves stable convergence within limited iterations and maintains high consistency under varying parameters, with identification precision and enforcement intensity as dominant factors influencing system stability.However, the current model is based on static networks and homogeneous agent assumptions. Complex heterogeneous structures and multi-layer semantic features are not fully captured, and deviations remain under extreme perturbations. Future work will incorporate dynamic graph modeling and multimodal feature fusion to enhance model adaptability and generalization.

References

[1] Hayashi Y, Tahmasbi N, Romanowich P, et al. Social and delay discounting in a behavioral economic analysis of bystanders' helping cyberbullying victims: The moderating role of gender[J]. Computers in Human Behavior, 2023, 145: 107783.

[2] Choudhuri A, Saraswat L. Explicating evolutionary epistemological concerns on gossip and cyberbullying[J]. Integrative Psychological and Behavioral Science, 2023, 57(4): 1331-1353.

[3] Lizhen O, Yiping Y, Jiao L, et al. An Agent-based Model of Opinion Dynamics with Hierarchical Thinking[C]//2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC). IEEE, 2024: 1330-1336.

[4] Wang R, Lu T, Zhang P, et al. Data-driven agent-based model for public opinion propagation simulation in cyberbullying[J]. Big Data Mining and Analytics, 2025, 8(4): 794-819.

[5] Gan C, Yang W, Zhu Q, et al. Hybrid rumor debunking in online social networks: A differential game approach[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2025, 55(4): 2513-2527.

[6] Su C, Xu Z, Wang X, et al. Research on the co-evolution mechanism of electricity market entities enabled by shared energy storage: A tripartite game perspective incorporating dynamic incentives/penalties and stochastic disturbances[J]. Systems, 2025, 13(9): 817.

[7] Philipo A G, Sarwatt D S, Ding J, et al. Cyberbullying detection: Exploring datasets, technologies, and approaches on social media platforms[J]. ACM Computing Surveys, 2026, 58(7): 1-35.

[8] Lekscha J, Mirbabaie M. How to analyze cyberbullying on social media platforms: A systematic literature review in information systems[J]. i-com, 2025, 24(2): 385-405.

[9] Meng X, Cai X, Li J. Combating cyberbullying with transparency: unveiling the impact of IP location disclosure on cyberbullying in Chinese social media[J]. Scientific Reports, 2025, 16(1): 174.

[10] Guo X, Jin J, Yan X. Online deviance in social media: a literature review and future research directions in the field of information systems[J]. Internet Research, 2025: 1-23.