Bots vs Guardians: Simulating Audit on Xiaohongshu (RedNote)

§1. Phenomenon Overview

Phenomenon

The rapid proliferation of automated accounts, or bots, in online platforms has significantly transformed digital interactions, raising concerns about misinformation, content manipulation, and the robustness of content moderation systems (Luo et al., 2023). Social media platforms face increasingly sophisticated threats from these automated bots, reshaping content visibility, engagement patterns, and user interactions. Xiaohongshu (RedNote), a hybrid e-commerce and social networking platform in China, exemplifies this vulnerability due to its dependence on personalized recommendation algorithms and community-driven content curation, making it highly susceptible to manipulative bot activities.

Ferrara's (2023) research highlights the growing sophistication of AI-driven bots, particularly their ability to mimic human behaviour and evade detection mechanisms (Ferrara, 2023). Our study explores the dynamic and adaptive interactions on Xiaohongshu. In this phenomenon, automated bots strategically manipulate engagement metrics, including likes, comments, and shares, to artificially inflate content visibility, creating deceptive trends and distorting genuine user interactions. In response, Guardian AI continuously refines its detection and moderation algorithms to identify and mitigate these evolving deceptive tactics effectively. This phenomenon essentially embodies a typical scenario of AI technology abuse: an escalating technological confrontation between malicious AI (bots) and defensive AI (Guardian AI), known as the "algorithmic arms race" (Pham et al., 2024).

Understanding this phenomenon is significant within media ecosystems. Firstly, the integrity of content ecosystems relies on users' trust in authentic engagement; artificially inflated metrics erode user trust and undermine platform credibility. Secondly, as platforms increasingly depend on algorithmic recommendation systems, manipulative bot behaviours can disproportionately amplify misinformation, damaging public discourse and potentially influencing societal outcomes, such as consumer behaviours and public opinions. Lastly, the escalating complexity of bot strategies necessitates equally adaptive moderation approaches, demanding continuous technological innovation and policy evolution within the digital media landscape.

Therefore, rigorous analysis of this adaptive AI-to-AI interaction is essential for developing robust strategies that ensure long-term platform integrity, protect user trust, and maintain a healthy digital ecosystem.

Problem statement

The widespread presence of bots has important implications for digital platforms and users. Zannettou et al. (2019) indicate that artificially inflated engagement metrics degrade users' trust and platform credibility. Furthermore, Gorwa et al. (2020) suggest that the reliance on algorithmic recommendation systems amplifies the negative impact of manipulative bot behaviours, disproportionately spreading misinformation and potentially influencing societal outcomes, such as consumer behaviours and public opinions. Additionally, Pham et al. (2024) emphasize that the sophistication and rapid adaptation of malicious bots present significant challenges for existing moderation frameworks, necessitating equally adaptive defensive strategies. Our study aims to deepen the understanding of these evolving interactions, highlighting critical vulnerabilities within current

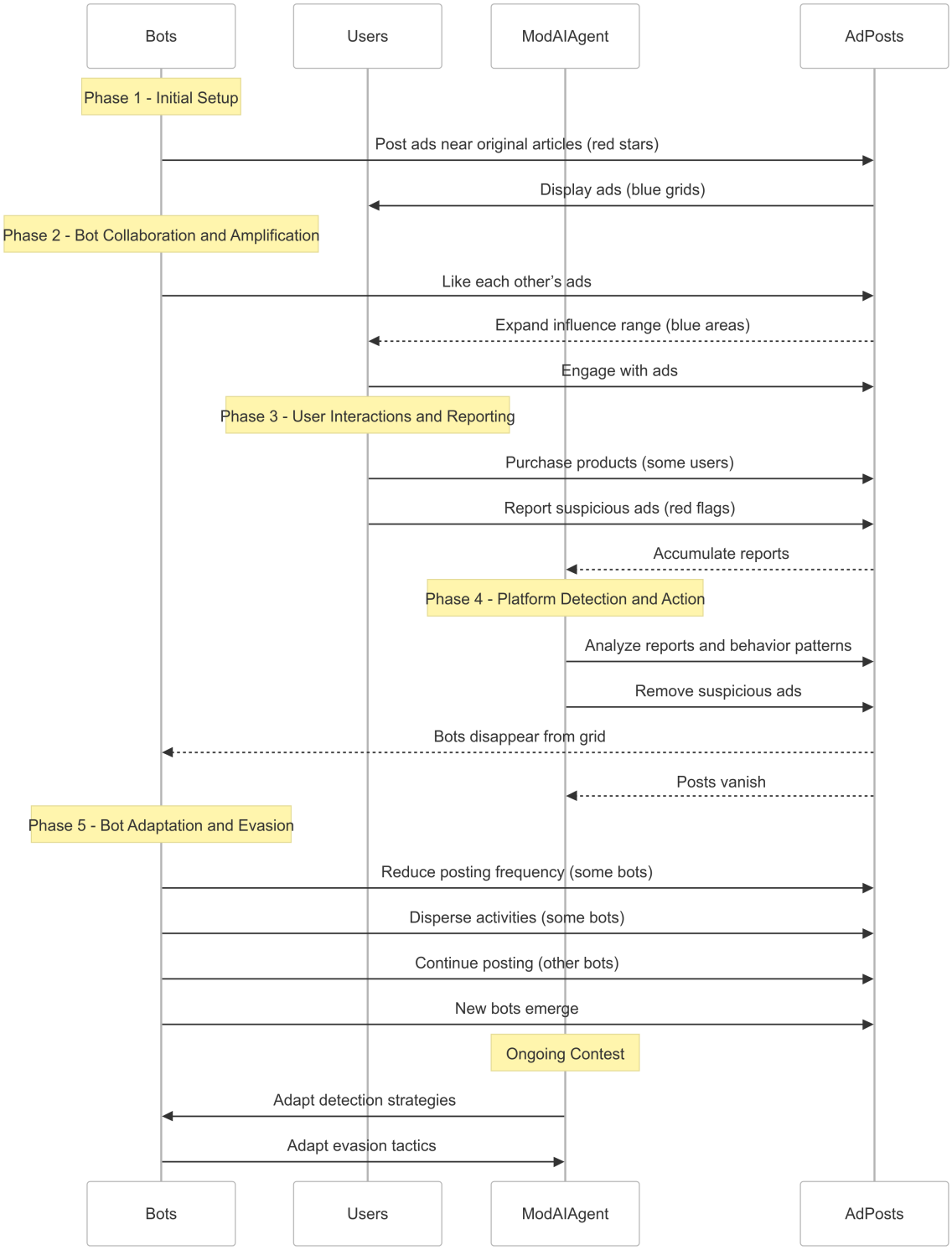
moderation systems and offering insights to inform the development of more robust detection mechanisms.

Suitability of Agent-Based Modeling

Agent-based modelling (ABM) is particularly suitable for studying this phenomenon due to its capacity to simulate complex adaptive systems composed of autonomous agents. ABM enables detailed observations of individual agent behaviours, their interactions, and the resulting emergent macro-level patterns. This methodological approach effectively captures the dynamic strategies and evolving interactions characteristic of AI-driven bots and Guardian AI, facilitating deeper insights into adaptive moderation and manipulative tactics.

Illustrate the Phenomenon

To "signpost" the content of this report, we present a curated sequence of visualizations from our agent-based simulation, capturing the emergence and progression of the phenomenon of interest: the interplay between ad/shill bots, human users, and a platform's detection AI in a media ecosystem. This sequence, described below as a series of snapshots at different timesteps, illustrates how bots collaborate to amplify ad content, how users respond, and how the platform's AI attempts to detect and mitigate their influence. Each phase includes text annotations to explain the progression over time, offering a narrative of this dynamic "cat-and-mouse" game.



- **Phase 1: Initial Setup**

The simulation begins with a 2D grid featuring original articles (represented as red stars) scattered across the space. A small number of ad/shill bots (green triangles) are introduced, each posting ad posts (blue grids) strategically near these articles to exploit their visibility and attract user attention.

- **Phase 2: Bot Collaboration and Amplification**
  As the simulation advances, the bots collaborate by liking each other's ad posts. This mutual liking increases the influence range of the posts, visualized as expanding blue areas around them. The amplified visibility draws more human users (yellow dots) to these ad posts, simulating how perceived popularity can enhance engagement in a media ecosystem.
- **Phase 3: User Interactions and Reporting**
  Human users begin interacting with the ad posts. Some, swayed by the high number of likes, purchase the advertised products, reflecting how bot-driven amplification can mislead users. Others, suspecting deception or fraud, report the posts, with these reports visualized as red flags accumulating on certain ad posts. This phase highlights the dual role of users as both potential victims and detectors.
- **Phase 4: Platform Detection and Action**
  The platform's detection AI monitors the grid, analyzing signals such as the number of reports and behavioral patterns. When a post accumulates sufficient reports or exhibits suspicious activity, the AI removes it or eliminates the bot responsible. Removed bots disappear from the grid, and their posts vanish, demonstrating the platform's efforts to curb malicious activity and restore ecosystem integrity.
- **Phase 5: Bot Adaptation and Evasion**
  Bots nearing detection—those with posts receiving multiple reports—adapt by reducing their activity, such as posting or liking less frequently, to evade the platform's thresholds. Meanwhile, other bots persist with their strategies, and new bots may emerge, perpetuating the cycle. This phase underscores the adaptive nature of bots in response to detection efforts.

This sequence of snapshots illustrates a continuous cycle of bot collaboration, user interaction, platform detection, and bot evasion. It reveals how ad/shill bots manipulate the media ecosystem through coordinated actions, how human users can be both deceived and instrumental in flagging suspicious content, and how the platform's AI strives to identify and remove malicious agents. The phenomenon emerges as an ongoing contest between AI-driven bots and the platform's detection mechanisms, with each side adapting to the other's strategies over time.

§2. Simulation Design & Implementation

System Overview:

The core components of the model include multiple agents interacting within a 2D grid-based environment to simulate the dynamics of ad dissemination and detection on social platforms. These agents are:

1. **Original Posts (Red Grids)**: Represent authentic content originating from users or brands.
2. **Ad Posts (Blue Grids)**: Represent advertisements posted by bots around original posts.
3. **Ad Bots (Green Triangles)**: Automated agents that post ads, collaborate with other bots, and attempt to evade detection.
4. **Human Users (Yellow Dots)**: Realistic user agents who interact with ad posts through actions such as liking, purchasing, or reporting.
5. **Platform Detection AI**: A system that monitors agent behavior, identifies suspicious accounts, and removes high-risk posts or bots.

The emergent phenomenon arises from the dynamic interactions among these agents. Ad bots

amplify their influence by collaborating to generate likes and comments, misleading human users into engaging with fraudulent content. The platform's detection AI counters this by analyzing abnormal patterns in engagement metrics and user reports, leading to an ongoing "algorithmic arms race" between manipulative bots and defensive AI systems.

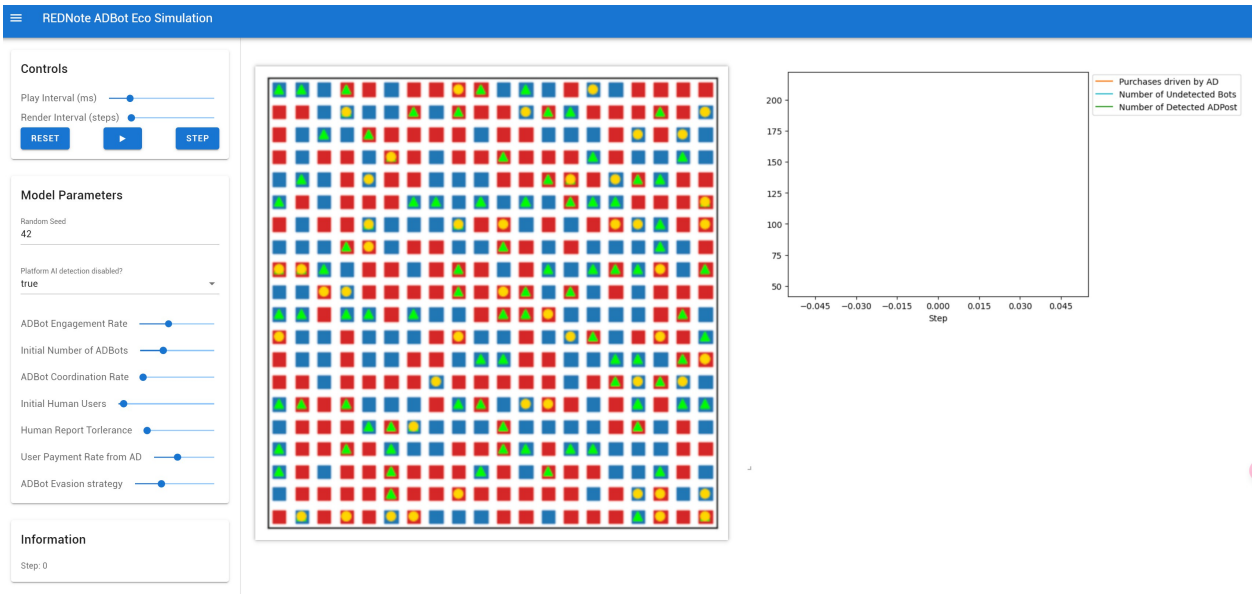


Figure 1: Simulation_Environment

Simulation Environment:

The simulation employs a hybrid spatial environment combining a **2D grid** for agent placement and movement with a **network structure** to represent social relationships. Key features of the environment include:

1. Mixed space environments
- GridSpace: GridSpace is a 40x40 two-dimensional grid where agents are randomly distributed in grid cells to simulate user and bot behaviors on a social platform by interacting with agents in neighboring cells. The grid simplifies location management and supports efficient visualization and presentation. Multiple agents may occupy the same grid cell, simulating the interactions of overlapping users and bots on the platform.
- Social Network (NetworkGrid): on top of the grid space, the model builds a network of social relationships to simulate the attention and interactions between users. Each user agent influences the propagation of content recommendations by following other users. The social network maps the behavioral patterns of users and their feedback on the recommended content.

2. Key agents and their behaviors
- Bot Agent: Bot agents publish fake advertisements and collaborate to increase exposure and bypass detection by audit systems. They adjust their behavioral strategies based on platform feedback to optimize the distribution of fake content.
- Audit AI Agent (ModAIAgent): Audit AI Agents are responsible for platform content detection, identifying fake ads through behavioral analysis and community detection, and dynamically adjusting their detection strategies in response to changes in bot behavior. The goal is to improve the quality of the platform content and maintain the integrity of the platform.

- HumanUserAgent: User agents interact with content based on the platform's recommendation system and influence the recommendation mechanism. Users can randomly report suspicious bots, and the interaction behavior in turn affects the recommendation system, which in turn affects the spread of false advertisements.
- Recommendation System (Recommendation System): The recommendation system dynamically adjusts the content weights according to user interactions (likes, comments, and reports), optimizes the recommendation through the feedback mechanism, and affects the dissemination of false advertisements and information flow.

3. Key parameters and constraints in the model
- Agent distribution: the initial position of agents in the grid space is randomized, which affects the frequency of interactions between agents. Social relationships form networks through attention and interaction, and user behavior affects recommender system optimization.
- Content exposure and interaction: robot agents increase the exposure of false advertisements through interaction. The recommendation system dynamically adjusts content weights based on user interactions, affecting the scope of false advertisement dissemination.
- Audit strategy: Audit AI agents respond to changes in robot behavior by dynamically adjusting their detection strategy, thus improving the quality of platform content.
- Social network feedback mechanism: social network connectivity and user behavior have an important impact on content dissemination, and users' attention behavior and social relationships affect the dissemination range of false advertisements.

Agent Design:

Ad Bot Behavior

- **Initial Strategy**: Ad bots seek out target original posts and publish ads nearby.
- **Collaborative Amplification**: Bots like each other's posts to artificially inflate popularity.
- **Evasion Tactics**: When facing high report risks, bots reduce posting frequency or disperse activities to avoid detection.

Platform Detection AI Design

- **Behavioral Analysis**: Monitors engagement patterns and flags abnormal growth rates or excessive reports.
- **Action Protocols**: Removes flagged posts or bans suspicious bots.

User Behavior

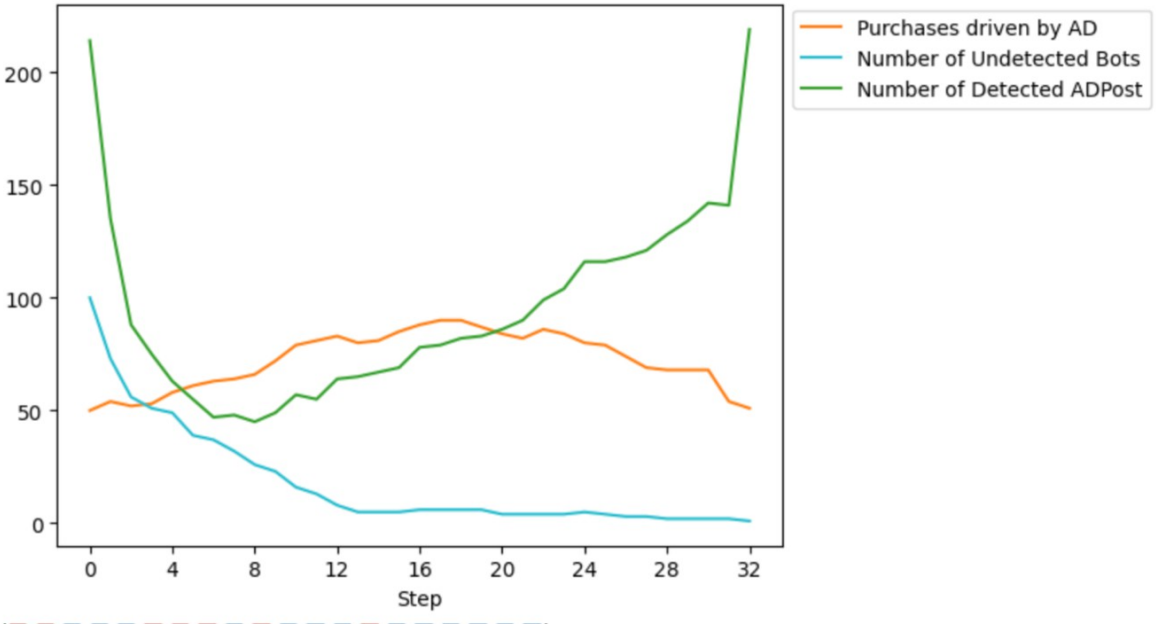
- Liking & Commenting: Influence the spread of content.
- Reporting: Increases the likelihood that the platform's AI will remove the content.
- Purchasing: Driven by exposure to high-ranking ads.
- For the computational implementation, we use an intelligent body model based on the MESA framework, where the bot and the platform AI interact in multiple time steps and adjust their strategies based on feedbacks

Interaction Dynamics:

- In the simulation, we use a combination of random scheduling and hierarchical scheduling:
  - Random scheduling: the bot randomly selects target users to publish advertisements to simulate the randomness of the bot in reality.

  - Hierarchical scheduling: the platform AI detects the content within a certain time interval, while the bot can quickly disseminate the information within the detection gap.

- This scheduling strategy captures the strategy game between the bot and the platform AI, allowing the bot to quickly adjust its strategy before the detection AI optimizes.

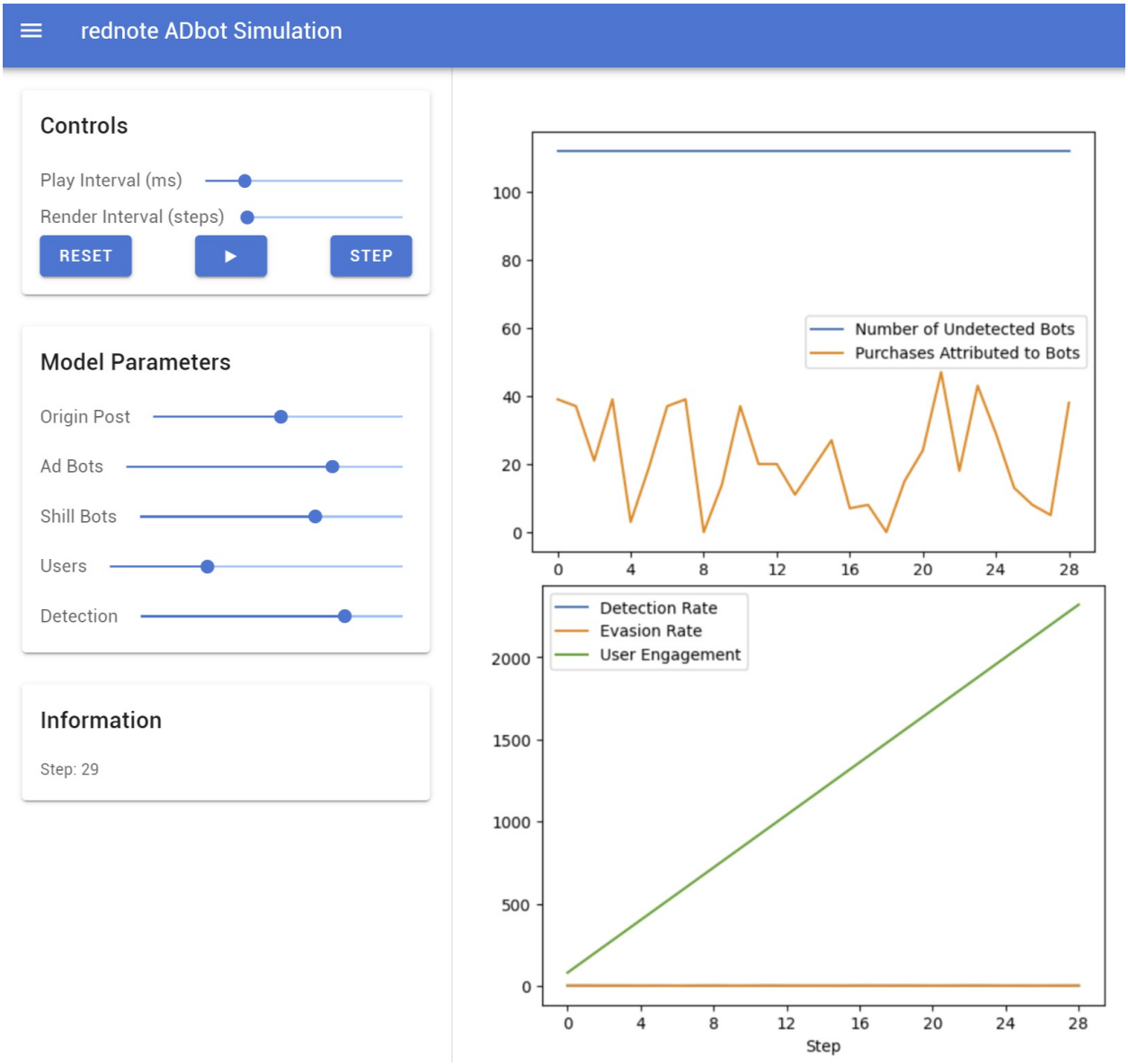


Data Collection and Visualization:

- The data we collected during the simulation include:
  - Bot Survival Rate: the percentage of bots that can still effectively disseminate information after a certain time step.

  - Platform AI detection success rate: the proportion of AI that successfully intercepts false content.

  - Purchase Volume: Number of purchases driven by bot-generated ads.



- Visualization methods include:

- **Grid Heatmaps**: Highlight areas of high interaction intensity.
- **Real-Time Statistics Panel**: Displays metrics such as undetected bots, purchase volume, and detection rate.
- **Curve Plots**: Track changes in key metrics over time.

Through these analyses, we can reveal the dynamic confrontation process between bot and platform AI, and provide optimization suggestions for the content regulation strategy of social media platforms.

§3. Observations & Results

Visualization of the simulation phenomenon

The results show that in the process of continuous strategy optimization, the interaction between the bot agent and the human user agent generates some complex emergent behaviors, which gradually

evolve from purely individual strategy adjustment to platform-level content dissemination and manipulation.

By running the simulation model several times and collecting data at different time steps, the following key aspects are observed and analyzed:

- Spreading rate of false advertisements: Initially, bots gain traction by posting ads and interacting with users. Over time, ModAIAgent improves detection, reducing ad exposure. The spread of false ads follows a non-linear trajectory, peaking before tapering off as detection mechanisms improve.
- Users exhibit diverse behaviors: Increased bot activity correlates with higher user reporting rates. Users gradually adapt to misinformation, altering their interaction patterns.
- Audit AI Detection and Response: ModAIAgent identifies bots using behavior analysis. Detection efficiency improves iteratively but fluctuates under coordinated bot attacks. Overreaction in detection can occasionally lead to false positives.

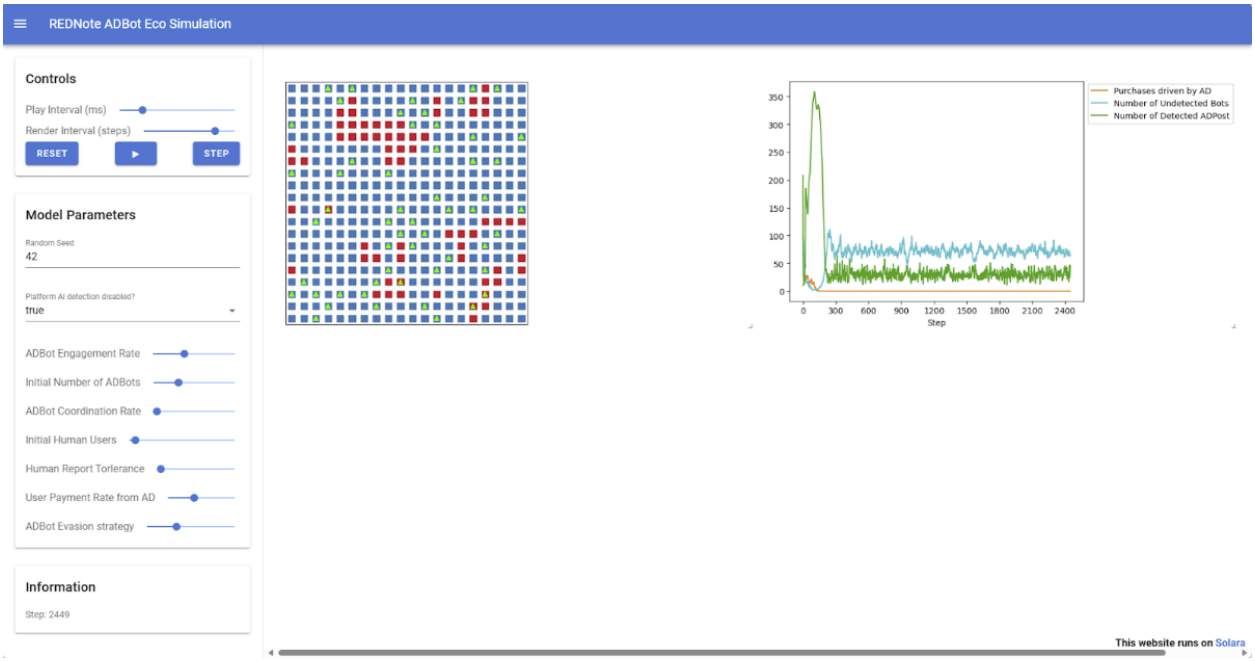


Figure 2: Dynamics

Quantitative Metrics and Qualitative Description

In order to analyze the simulation results more systematically, we use several key metrics to measure the dissemination effect of false advertisements as well as the interaction patterns between bot agents and user agents.

- Exposure rate of false ads: We measure the spreading effect of false advertisements by recording the exposure frequency of each advertisement on the platform. Figure 1 illustrates the trend of the exposure rate of false ads over time. It can be seen that the exposure rate is low initially, but increases significantly as the frequency of advertisements posted by bot agents increases. The intervention of auditing AI makes the exposure rate gradually fall back at the later stage.
- Frequency of User Reporting: The reporting behavior of user agents on false advertisements is an important indicator of the platform interaction mode. Figure 2 shows the reporting behavior of users at different time steps. The data shows that as the number of false ads increases, the

frequency of user reporting rises rapidly, indicating that users will gradually take proactive actions in the face of false content.
- Audit AI Detection Rate: The Audit AI agent evaluates its efficiency by the success rate of detecting false advertisements. Figure 3 demonstrates the ability of Audit AI to recognize false ads at different stages. Although the robot agent's strategy makes the Audit AI's detection efficiency low at the beginning, the Audit AI gradually optimizes its detection strategy over time, and eventually achieves a high detection success rate.

Unexpected Behavior and Emergence Dynamics

Although we anticipated during model design that the robotic agents would gradually adjust their strategies as the detection mechanisms of the auditing AI were optimized, we observed some unexpected behavioral patterns.

- Collaborative Behavior of Robotic Agents: Over the course of multiple simulation runs, we noticed a gradual increase in collaborative behavior among the robot agents. Certain bots started to increase their exposure by co-publishing similar fake advertisements, and even manipulating the social network on the platform by liking and commenting on each other's ads. This collaborative behavior creates an advertising "information bubble" that dramatically increases the impact of false advertisements in a short period of time, even though the auditing AI agents have begun to strengthen their detection.

- User Agents' Response Patterns: User agents exhibit complex response patterns when confronted with fake ads. Some users were not initially suspicious of the ads, but their reporting behavior increased significantly as more false ads appeared. However, certain users consistently failed to recognize all false ads, possibly due to the diversity of ad content on the platform and the varying interests of users. This phenomenon suggests that the diversity of content in recommender systems and platforms may have an impact on users' ability to recognize false ads.

- Overreaction of Audit AI: Although the detection ability of the Audit AI is gradually improving, in some cases the Audit AI exhibits overreaction, especially when faced with collaborative interactions with bots. The Audit AI's detection mechanism sometimes misjudged the interaction behavior as false, thus incorrectly flagging certain legitimate advertisements. This misjudgment phenomenon suggests that Audit AI needs more refined detection strategies to avoid overreaction.

Interpretation and Discussion

Collaborative Behavior of Bot Agents: By analyzing the collaborative behavior of bot agents, we find that the spread of false advertisements is not only driven by individual bots, but also by mutual collaboration among multiple bot agents. This finding reminds us that the spread of false advertisements is no longer just an interaction between individual users and advertisements, but can spread rapidly through the collaboration of bots, forming a large-scale information bubble.

Diversity of user responses: User responses to false ads are complicated by personalized recommendation systems. While the platform's recommendation algorithms provide users with accurate ad recommendations, they may also lead to user bias when confronted with false ads. Therefore, the content diversity of the platform's recommendation system directly affects the ability of users to recognize false advertisements. We suggest strengthening the support for advertisement recognition when designing recommendation algorithms in the future to reduce the misleading of users by false

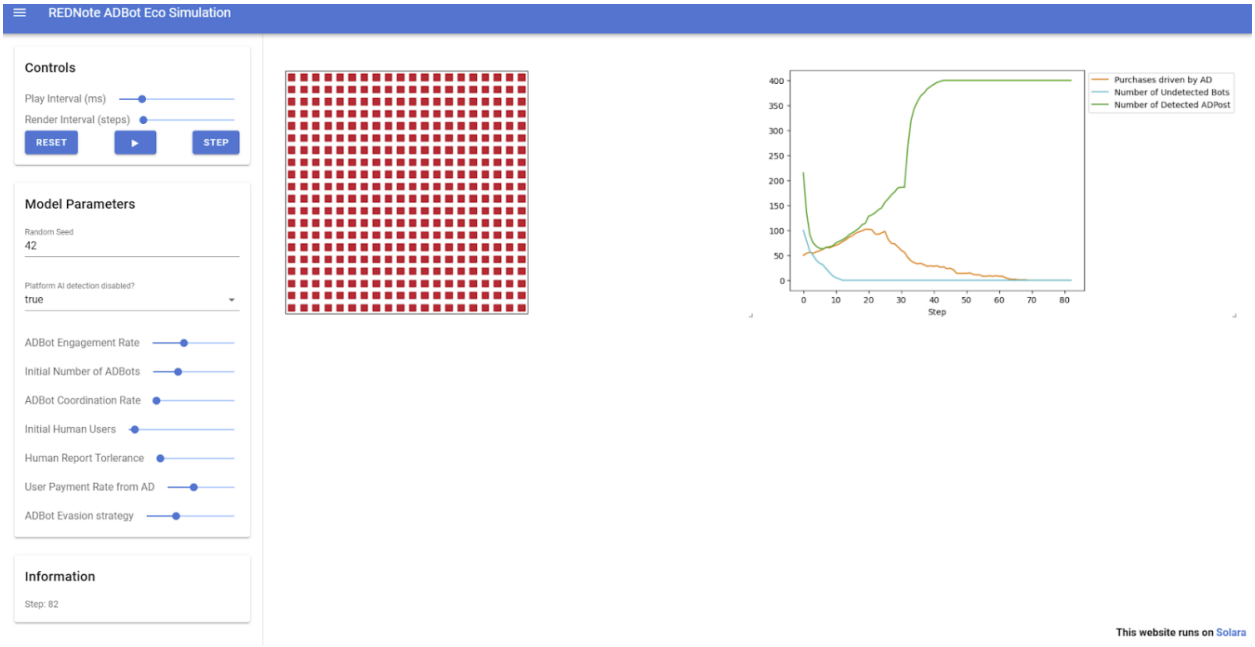


Figure 3: UnexpectedBehavior

advertisements.

Improvement suggestions for auditing AI: According to the results in the simulation, the misjudgment problem of auditing AI when facing the collaborative behavior of bots suggests that our auditing mechanism needs to be further optimized. Especially when facing the complex advertising network effect, the AI system should be able to identify the collaborative behaviors in the advertisement dissemination and perform more fine-grained detection in order to avoid misjudging normal advertisements.

Conclusion

Through the simulation analysis in this chapter, we find that the collaborative behavior of bot agents significantly promotes the dissemination of false advertisements and breaks through the platform's audit mechanism. In addition, the platform's recommendation system may also mislead users and fail to identify false advertisements effectively, highlighting the challenges of personalized recommendation systems in detecting the authenticity of advertisements.

For the auditing AI, we observed that it overreacted and misjudged in the face of robot agent collaboration, affecting the normal dissemination of advertisements. Therefore, improving content auditing mechanisms and recommender systems is a key future direction to more effectively address the challenges of false advertisements.

§4. Ethical & Societal Reflections

Ethical Considerations

The increasing reliance on AI-driven moderation systems raises critical ethical concerns, particularly around bias, privacy, transparency, and potential unintended consequences of automated decision-making processes (Hakami et al., 2024). Although our project utilized synthetic rather than real-world data to protect individual privacy, the simulation still prompts important ethical

considerations regarding data usage. Specifically, if such a simulation were applied or adapted to real social media platforms, there would be significant privacy implications. Bots could potentially impose user interaction data such as likes, comments, and sharing patterns to optimize the dissemination of false advertisements, leading to intrusive surveillance without explicit user consent. Therefore, transparent communication from platforms regarding the collection, usage, and protection of user data is paramount to maintaining user trust and ethical compliance.

Moreover, automated moderation tools, despite their efficiency, may unintentionally amplify biases embedded within algorithmic training datasets (Gorwa et al., 2020). For instance, AI-driven moderation systems risk disproportionately silencing certain groups or viewpoints due to biased historical moderation data. These concerns underline the necessity of ongoing ethical oversight and governance frameworks to prevent such biases from proliferating. Platforms must remain vigilant in continuously auditing their moderation algorithms, ensuring fairness, accountability, and transparency in automated decision-making processes.

Societal Implications

Our simulation reveals critical insights into the entangled dynamics of bot-to-bot and human-bot interactions, with societal consequences unfolding across micro, meso, and macro levels. These insights reflect both the potential and perils of AI deployment in content ecosystems, particularly in platforms like Xiaohongshu (RedNote), where social influence is algorithmically curated and amplified.

- **Micro-level:**

At the individual level, users are particularly vulnerable to subtle manipulations when bots artificially inflate likes, comments, and engagement signals. These surface-level indicators often influence user decision-making, such as whether to engage with or trust a piece of content. When bots collude to amplify ad posts, real users may unknowingly participate in the spread of misinformation or deceptive marketing. This mirrors real-world incidents where users have been misled by inauthentic reviews or engagement metrics, eroding trust not only in specific content but also in the platform as a whole.

Our simulation shows that even when users eventually report suspicious content, the delay in detection allows false content to gain traction. This parallels studies showing that misinformation spreads more quickly than corrections (Vosoughi et al., 2018), reinforcing the idea that early exposure has disproportionate influence.

- **Meso-level:**

At the community level, the presence of coordinated bots disrupts the credibility of recommendation systems and weakens social cohesion. In our simulation, once bot behaviour becomes dominant within specific regions of the platform, user reports become less frequent due to the normalization of manipulative content, which is an emergent behaviour reflecting trust fatigue.

This aligns with observed phenomena on platforms like Facebook or Twitter, where persistent exposure to low-quality or misleading content has been linked to user disengagement, increased polarization, and platform abandonment. When moderation systems are perceived as ineffective or biased, users begin to question the integrity of the entire ecosystem. Our model demonstrates this decline through diminishing report frequency and stagnant improvement in AI detection effectiveness without timely user feedback.

- **Macro level:**

Societal Stability and Governance ChallengesAt the societal level, bot-driven misinformation campaigns can severely disrupt democratic processes, polarize societies, and distort critical public discourses. The significant influence of misinformation in events like the 2016 U.S. presidential elections and misinformation surrounding the COVID-19 pandemic highlights the broad societal risks posed by ineffective moderation systems. These real-world scenarios underscore the urgent need for comprehensive, interdisciplinary collaboration between technology companies, policymakers, researchers, and civil society to effectively govern digital ecosystems and protect societal stability.

Potential for malicious use

Although this model is designed to explore the mechanisms of false advertisement propagation, it may also be used for malicious purposes. Malicious users could potentially exploit simulation tools like ours to develop more evasive and strategic bot behaviours for spreading misinformation and manipulating public opinion. For example, by analyzing detection weaknesses or exploiting behavioural feedback loops, adversaries could engineer bots capable of avoiding detection while maintaining strong influence.

To mitigate this risk, platforms must proactively invest in stronger content auditing systems, advanced bot detection mechanisms, and continuous research into evolving adversarial tactics. One of the inherent risks in publishing bot behaviour simulations is their dual-use nature: while they offer insight for defenders, they can also inform attackers. Therefore, ethical dissemination of findings, selective sharing of technical details, and robust collaborations between researchers, platform administrators, and ethical review boards are essential to prevent misuse and ensure these tools serve the public good.

Recommendations for Ethical AI Governance

To comprehensively address these ethical challenges and societal impacts, we propose actionable recommendations anchored in broader AI governance frameworks:

Enhanced Transparency: Platforms should transparently disclose all data collection practices, provide clear mechanisms for user consent, and enable users to control their personal data actively.

Continuous Bias Audits: Regularly auditing moderation algorithms to detect and address inherent biases is vital to ensuring fairness, inclusivity, and ethical integrity within digital environments.

Interdisciplinary Collaboration: Active partnerships among technologists, policymakers, ethicists, and social scientists can foster robust, adaptable governance frameworks that effectively respond to evolving threats posed by bot-driven misinformation.

Public Digital Literacy Initiatives: Expanding education programs on digital literacy will empower users to recognize misinformation, understand algorithmic manipulation, and engage proactively in content moderation.

By actively incorporating these recommendations, platforms can ensure ethical integrity, strengthen resilience, and uphold societal trust, effectively mitigating the adverse impacts of automated misinformation and bot interactions.

§5. Lessons Learned & Future Directions

Design and Development Reflections:

During the design and development of the simulation, the team encountered several complex and interdependent challenges, particularly in the areas of agent behaviour modelling, simulation realism, and network dynamics. Initially, the behavioural logic of bot agents was relatively static and deterministic, failing to capture the adaptive and coordinated nature of malicious bots observed on real platforms. Bots were initially configured to post ads without accounting for contextual interactions, environmental feedback, or evolving strategies based on detection.

As the simulation matured, we identified the need to simulate more realistic behaviours, such as feedback-driven adaptation, social engineering tactics, and cross-bot collaboration. We addressed this by introducing a feedback mechanism allowing bots to alter their tactics based on user interactions—such as likes, comments, and reports—which enabled more authentic patterns of spread and evasion to emerge.

In parallel, the social network layer underwent progressive refinement. We transitioned from a simplistic uniform user graph to a heterogeneous, dynamically responsive network. This revised design allowed us to model varying degrees of user influence, localized clustering, and the formation of social echo chambers—critical for studying real-world manipulation dynamics.

Another major milestone was the integration of a user behavior-aware recommendation engine. Initially based on random ad delivery, this mechanism evolved into a reactive recommendation system that adjusts ad visibility according to user interaction history. This update allowed us to more closely simulate algorithmic amplification's influence, a major vulnerability exploited by bots in real-world systems like Xiaohongshu.

Model Limitations & Areas for Improvement:

While our simulation provides useful insights into the dynamics of bot-driven content manipulation, several limitations constrain its fidelity to real-world media ecosystems.

One key limitation lies in the bot agents' lack of genuine learning capabilities. Although our bots can adjust behaviours based on platform feedback (such as detection or user reports), their responses follow predefined rules rather than adaptive learning. In real-world scenarios, adversarial bots increasingly employ reinforcement learning and large language models (LLMs) to evolve context-sensitive strategies. For example, bots on platforms like Twitter or Xiaohongshu can vary the linguistic style, adapt to regional dialects, and alter behaviour based on real-time sentiment trends. Future simulation iterations should incorporate reinforcement learning frameworks and NLP-powered text generators to better reflect these dynamics, allowing bots to refine tactics across simulation cycles.

A second limitation relates to the simplicity of the content propagated by bots. In the current model, false advertisements are relatively uniform and easy to detect. However, deceptive content in actual ecosystems often appears in subtle forms, such as fake testimonials, simulated peer reviews, or contextually embedded misinformation. These forms are difficult to detect with keyword-based rules alone and often evade moderation tools. Expanding the model to include synthetically generated review content, visual cues, or multimodal deception strategies would significantly improve its realism and detection testing capabilities.

The third limitation involves how the simulation represents social network structures and agent heterogeneity. While our model includes a basic social graph layered on top of a 2D grid, it does not yet fully capture real-world social dynamics such as echo chambers, clustered opinion groups, or influencer-driven dissemination. Nor does it account for user-level diversity in susceptibility to deception, cognitive biases, or reporting behaviours. Enhancing the network topology to include

modular community detection, preferential attachment mechanisms, and agent-specific traits (e.g., trust thresholds or political leaning) would allow more accurate modelling of how content propagates and amplifies within segmented digital publics.

Finally, computational scalability presents a constraint on model complexity and experimentation. As agent count and interaction logic increase, the simulation experiences performance bottlenecks in scheduling and data logging. While the Mesa framework offers flexibility, future versions should consider parallelized execution or agent batch processing, potentially via multiprocessing or GPU acceleration. Integrating modular logging systems or offloading visualization tasks could further optimize runtime performance, enabling the simulation of larger-scale ecosystems with finer behavioural resolution.

Addressing these limitations will not only improve simulation realism but also broaden its applicability for testing real-world moderation interventions and AI safety strategies.

Future Applications:

The insights derived from our simulation offer meaningful contributions to platform governance, AI safety research, and digital policy development. From a governance perspective, this framework enables social media platforms such as Xiaohongshu to understand better how bot strategies exploit recommendation systems and user trust. By simulating both evolution and detection gaps, platforms can preemptively evaluate moderation strategies and refine auditing algorithms in response to emerging tactics.

In the realm of AI safety, our findings reinforce the importance of adaptive moderation tools in adversarial digital environments. As Bradshaw et al. (2021) emphasize, ensuring the robustness of automated systems requires ongoing recalibration in response to intelligent threats. Our simulation provides a controlled testbed for modelling these adaptive behaviours, allowing researchers to evaluate how detection algorithms perform under shifting conditions. This could inform the design of resilient content moderation systems better equipped to manage misinformation, coordinated inauthentic behaviour, and adversarial AI.

On a policy-making level, the simulation's outputs can help regulatory bodies assess moderation interventions' efficacy and potential unintended consequences. For example, it allows policymakers to explore trade-offs between overly aggressive takedown thresholds (risking over-censorship) and lenient moderation (permitting disinformation spread). These simulated outcomes can provide evidence for designing more transparent, accountable, and flexible moderation protocols aligned with digital rights and democratic values.

Furthermore, our model can serve as an educational and strategic planning tool. Civil society organizations, academic institutions, and technology ethics boards can use it to understand better the propagation of false content, train platform auditors, or simulate crisis scenarios. By providing concrete data and dynamic visualizations, the simulation supports multidisciplinary collaboration among technologists, social scientists, and policymakers in building safer and more trustworthy digital ecosystems.

§6. References

- Bradshaw, S., Neudert, L. M., & Howard, P. N. (2021). Industrialized disinformation: 2020 global inventory of organized social media manipulation. Oxford Internet Institute, University of Oxford. https://doi.org/10.5281/zenodo.4615251*

- *Ferrara, Emilio. (2023). Social bot detection in the age of ChatGPT: Challenges and opportunities. First Monday. 10.5210/fm.v28i6.13185.*
- *Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data & Society, 7(1). https://doi.org/10.1177/2053951719897945*

- *Hakami, Ammar & Tazel, Ravi. (2024). The Ethics of AI in Content Moderation: Balancing Privacy, Free Speech, and Algorithmic Control. 10.13140/RG.2.2.19529.97121.*

- *Luo, H., Meng, X., Zhao, Y., & Cai, M. (2023). Rise of social bots: The impact of Social Bots on public opinion dynamics in Public Health Emergencies from an information ecology perspective. Telematics and Informatics, 85, 102051. https://doi.org/10.1016/j.tele.2023.102051*

- *Pham, B. C., & Davies, S. R. (2024). What problems is the AI act solving? Technological solutionism, fundamental rights, and trustworthiness in European AI policy. Critical Policy Studies, 1–19. https://doi.org/10.1080/19460171.2024.2373786*

- *Raees, M., Meijerink, I., Lykourentzou, I., Khan, V.-J., & Papangelis, K. (2024). From explainable to interactive AI: A literature review on current trends in human-ai interaction. International Journal of Human-Computer Studies, 189, 103301. https://doi.org/10.1016/j.ijhcs.2024.103301*
- Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science, 359(6380), 1146–1151. https://doi.org/10.1126/science.aap9559*
- *Zannettou, S., Sirivianos, M., Blackburn, J., & Kourtellis, N. (2019, January 18). The web of false information: Rumors, fake news, hoaxes, clickbait, and various other shenanigans. arXiv.org. https://arxiv.org/abs/1804.03461*

§7. Attestation

This report reflects the collective effort of all group members. Below is a summary of each member's contributions based on the CRediT Contributor Role Taxonomy.
1. Jiayi Chen (Jaye) - Data Analysis and Report Draft Writing: Conducted data collection and analysis.
2. Xintong Ling (Sylvia) - Conceptualization, Project administration, Writing – review & editing, Supervision.
3. Huanrui Cao (Saikoro) - Report Review and Model Improvement: Reviewed and validated the report for accuracy and clarity.