## §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). Recent studies highlight the growing sophistication of AI-driven bots, particularly their ability to mimic human behavior and evade detection mechanisms (Ferrara, 2023). These bots interfere with the platform's ecology through keyword triggering, collaborative dialog faking, and dynamic detection evasion.On platforms such as RedNote, bots are faking fake buzz by manipulating likes and comments to boost the exposure of advertised posts. This phenomenon essentially reflects a typical scenario of AI technology abuse: the escalating technological confrontation between malicious AI (bots) and defensive AI (Guardian AI), the so-called "algorithmic arms race" (Pham et al., 2024).Our project investigates the dynamics of AI-to-AI interactions within the context of content moderation, focusing on the strategies employed by automated systems to counteract malicious bot activity (Raees et al., 2024).

- **Problem statement:**

- 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). Recent studies highlight the growing sophistication of AI-driven bots, particularly their ability to mimic human behavior and evade detection mechanisms (Ferrara., 2023). This phenomenon essentially reflects a typical scenario of AI technology abuse: the escalating technological confrontation between malicious AI (bots) and defensive AI (Guardian AI), the so-called "algorithmic arms race" (Pham et al. 2024). Our project investigates the dynamics of AI-to-AI interactions within the context of content moderation, focusing on the strategies employed by automated systems to counteract malicious bot activity (Raees et al., 2024). To illustrate this phenomenon, we have curated a sequence of visual representations that depict the evolution of bot activities and moderation strategies over time. These visualizations help demonstrate how content moderation algorithms adapt to evolving bot tactics and how emergent patterns of manipulation unfold within a controlled simulation environment. Through this approach, our study aims to contribute to a deeper understanding of the challenges facing AI-driven moderation efforts and the potential gaps in existing automated detection systems (Zannettou et al., 2019).

- **Why agent-based modeling**

- The Intelligentsia-based model (with MESA as the implementation framework) was chosen as an analytical tool for this phenomenon because of its ability to automate the simulation of different types of interactions and strategies and to test their effects. At the same time, macro-phenomena (e.g., a decrease in the overall credibility of the platform) arise naturally from micro-interactions (e.g., the commenting behavior of a single bot). In addition, the intelligible

model demonstrates how automated strategies act on interaction cycles and platform media structure generation.

- **Illustrate the Phenomenon**:

- Description of the phenomenon:
  The following sequence shows the evolution of the bots and the platform AI as they interact:
  - Phase 1: Bots start posting advertising messages in large numbers, some of which use deceptive headlines and highly similar user-generated content.
  - Phase 2: The platform AI detection system identifies and removes some of the advertising comments, but some bots manage to bypass detection and continue to propagate by adjusting the text content or embedding unstructured data.
  - Stage 3: Bots employ more advanced strategies, such as using deep learning to generate more natural-looking ad content, making it harder to distinguish.
  - Stage 4: Platform AI introduces new detection mechanisms, such as using graph neural networks (GNNs) to analyze anomalous patterns in social networks and improve detection accuracy.
  - Stage 5: The confrontation between bots and platform AI enters a new phase, where platform AI optimizes its algorithms to adapt to changing bot strategies, while bots attempt to leverage user interaction data to simulate normal user behavior.

- This evolution clearly demonstrates the trend of AI confrontation in the social media ecosystem and the need for future regulatory strategies to constantly adapt to this dynamic change in order to effectively curb the proliferation of disinformation.

## §2. Simulation Design & Implementation

- **System Overview**:

- The central goal of this research is to model the confrontation between automated bots and platform AI on social media platforms, aiming to study their dynamic behavior in the information ecosystem. The dynamic confrontation in the model is reflected in the fact that the bots constantly adjust their strategies to bypass detection, while the platform AI optimizes its algorithms to improve recognition. The emergence of this phenomenon mimics the challenges of content censorship in reality. The core components of the system include three main agents: the Bot Agent, the Review AI Agent (ModAIAgent), and the Real User Agent (HumanUserAgent). In addition, the recommender system dynamically adjusts the visibility of the content, which in turn affects the interaction and information dissemination among the agents. The simulation employs the Mesa framework for agent modeling and visualization; the spatial structure of the model uses GridSpace for compatibility with Mesa's visualization tools, and simulates the geographic location of agents and their interaction patterns in social platforms.

- **Simulation Environment**:

- This simulation model is based on a hybrid spatial environment combining GridSpace and NetworkGrid for simulating interactions between robotic agents, real user agents and auditing AI agents. The environment simulates a complex social platform where the behavior of the agents is not only affected by their physical location but also by the dynamic changes in social relationships.
  1. Mixed space environments

- GridSpace: GridSpace is a 50x50 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**:

- Bot Design
  - Initial strategy: the bot will use a simple template to generate ad content.

  - Evolutionary mechanism: When a bot's ad is removed by the platform AI, it will adjust its strategy, including changing the text structure, adding random elements, or even mimicking user behavior.

  - Social engineering: some bots may mimic the interaction patterns of real users to increase survival rates.
- Platform Detection AI Design
  - Rule-based detection: match known ad patterns.

  - Machine learning detection: adjusting classification models based on training data.

  - Behavioral analysis: detect abnormal interaction patterns, such as an account posting a large number of similar comments in a short period of time.
- 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.
  - User reporting behavior: the frequency of user reporting and its impact on the detection success rate.

- Visualization methods include:
  - Network structure diagram: show the interaction between bot and users.
  - Time series graph: tracking the change of bot adaptation strategies.
  - Heatmaps: to show the distribution of detection efficiency of platform AI.

- 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.

- **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

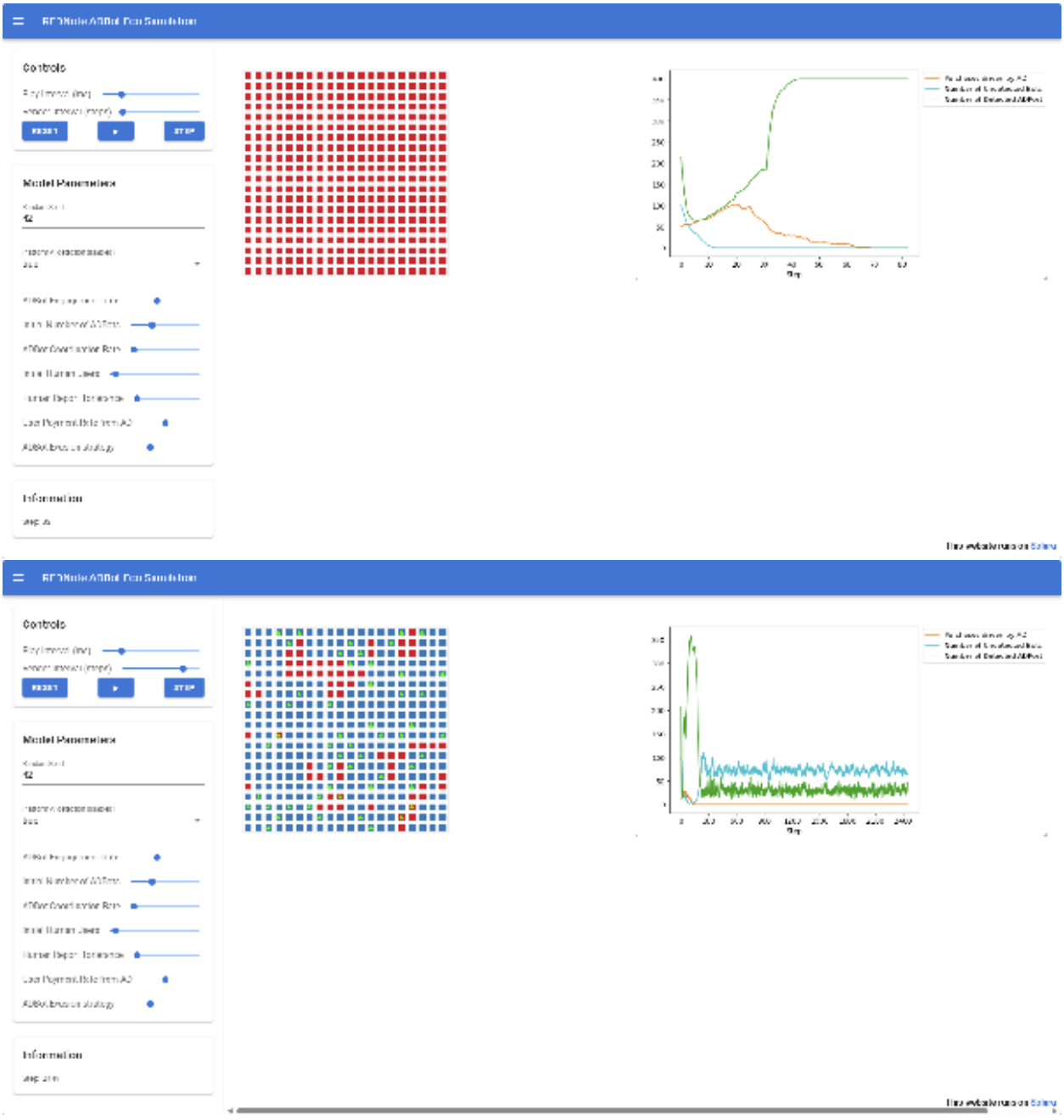
beginning, the Audit AI gradually optimizes its detection strategy over time, and eventually achieves a high detection success rate.

- **Charts and Network Diagrams**

- Visual Representations

- Red nodes: Represent original content.

- Blue nodes: Indicate advertisement posts.



- **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 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 raises critical ethical concerns, particularly regarding bias, privacy, and the transparency of automated decisions (Hakami et al., 2024). Our project primarily focuses on synthetic data to avoid direct privacy issues, yet the broader implications of real-world AI moderation must be considered. Prior research has demonstrated that automated moderation systems can reinforce existing biases in content filtering and suppression (Gorwa, 2020). In this project, we do not use real social media data, but simulated data and behaviors to study the propagation of false advertisements and their interaction with users. The agent behaviors in the model are entirely based on hypothetical rules and do not involve any real users' privacy data.

- Despite this, privacy risks could still arise if this model were applied to a real platform. For example, in a scenario where a robotic agent optimizes advertisement dissemination by collecting user interaction data (e.g., likes, comments, shares), privacy issues could emerge. Users might not be aware that their interactions are being used to improve an advertisement's targeting, potentially violating their privacy, especially if such data collection occurs without explicit consent. To address these concerns, platforms must implement strict privacy protection measures and engage in ethical scrutiny regarding the collection and use of advertising data. Transparency in how data is collected, stored, and used is essential to ensuring user trust and protecting privacy in an AI-driven content ecosystem.

- **Societal Implications**:

- The implications of false advertisement propagation on social platforms are multi-faceted, affecting individuals, communities, and society at large. Our findings show that these issues play out on several levels.
  - **Micro-level:**

  - At the individual level, bot advertisements have the potential to influence user behavior before they even realize the content is false. For example, users may engage with an ad without questioning its legitimacy, leading to a temporary shift in their perceptions or actions. Even if users can eventually report and remove the false advertisement, the initial exposure could have already decreased their trust in the platform or the content they interact with. This could result in a lingering sense of distrust or skepticism among users, undermining the overall user experience.

  - **Mesoscopic level:**

  - At the community level, the effectiveness of social platform AI systems—particularly recommender and audit algorithms—can significantly impact the spread of false advertisements. When platforms fail to identify and limit the propagation of false ads, the trust in the platform itself is jeopardized. A failure in this regard could decrease user engagement, as users may begin to question the platform's credibility. This decline in user trust could lead to a deterioration in platform health, as users become less inclined to participate in the platform's ecosystem or rely on its content for information.

  - **Macro level:**

  - At the broader societal level, the spread of false advertisements can have far-reaching effects on public opinion and social dynamics. False advertisements may influence voting behavior during elections, sway public sentiment on critical issues, or even cause social unrest. These kinds of manipulations can disrupt democratic processes, leading to significant consequences for public discourse and societal stability. Therefore, effective content moderation is not just about maintaining platform integrity but also about safeguarding broader societal structures and preventing misinformation from distorting reality on a mass scale.

- **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 may use the simulation tool to design more sophisticated false advertising strategies that manipulate user behavior. To prevent this, platforms need to enhance content auditing and bot behavior detection to ensure that these techniques are not misused.
  One of the inherent risks of creating simulation tools for understanding bot behavior is the potential for malicious use. Although the goal of this model is to explore the propagation of false advertisements and their interaction with users, malicious users may exploit similar tools to design more sophisticated strategies for manipulating online platforms.
  To mitigate this risk, platforms must invest in stronger content auditing systems, advanced bot

detection mechanisms, and ongoing research into evolving strategies for combating malicious behaviors. Collaboration between researchers, platform administrators, and ethical bodies will be essential in ensuring that such tools are not misused for harmful purposes.

## §5. Lessons Learned & Future Directions

- **Design and Development Reflections**:

- During the simulation design and implementation process, the team faced several challenges, especially in intelligent body behavior modeling and social network simulation. Initially, the behavioral model of the robotic intelligences was relatively simple and could not adequately simulate the complex social platform dynamics. The intelligences simply posted advertisements without considering the interactions between users and the propagation paths of advertisements in social platforms. As the simulation development progressed, the team realized the need for a more refined behavioral model of the intelligences, including social engineering strategies, feedback mechanisms, and collaborative behaviors across intelligences.

- To address these challenges, the team continuously adapted the behavioral rules and interaction strategies of the intelligences. Instead of just choosing actions based on random selection, the intelligences introduced a feedback mechanism that allowed them to adjust and optimize based on user feedback (e.g., clicking, commenting, reporting, etc.), thus simulating the spreading process of false advertisements more realistically. At the same time, the design of the social network has undergone several optimizations, with the original simple user relationship gradually transformed into a more complex social network structure. In the new design, the interaction between users is more dynamic, and users are not only simple "friends", but also have richer levels of influence, which can better simulate the dissemination of information and interaction on social platforms.

- In addition, the recommendation mechanism of the simulation system has also undergone a critical adjustment. The original simple ad push mechanism is upgraded to a recommendation engine based on user behavior. Users' interactions with advertisements (such as clicks, likes, and comments) will directly affect the display frequency and visibility of advertisements, thus more accurately simulating the effect of the platform's algorithmic recommendation in reality.

- **Model Limitations & Areas for Improvement**:

- Although current simulation models provide valuable insights into the spread of false advertising, there are still some notable limitations that restrict the fidelity of the models to real-world phenomena.

- First, robotic intelligences lack sufficient learning capabilities and adaptive mechanisms. Although the intelligences are able to make adjustments based on the platform's detection mechanisms, their behaviors do not have true learning capabilities. In reality, the automated system behind false advertisement dissemination is based on deep learning and adaptive algorithms, which enables the bots to adjust their strategies in real time in a constantly changing social environment. In the future, more sophisticated learning algorithms should

be introduced to enable bots to continuously improve their communication efficiency and strategy complexity by simulating the human learning process.

- Second, the current false advertising content is relatively simple. False advertisements in the real world are not only obvious fraudulent advertisements, but also include some subtle and hard-to-detect forms of advertisements, such as fake product reviews and false social proofs. These advertisements are usually highly covert and difficult to recognize through traditional rule detection mechanisms. Therefore, future improvements could include a greater variety and complexity of advertisement forms to better mimic real-world false advertisement behaviors and improve the detection accuracy of the simulation model.

- In addition, the design of social networks remains an area for improvement. Although current models consider interactions between users, the diversity of information dissemination mechanisms and user behaviors in social networks is still not fully reflected. The complexity and variability of user behaviors in real platforms, including factors such as viral dissemination of information, the influence of opinion leaders, and localized dissemination of information, are all missing in the current model. Therefore, the complexity of social networks should be further enhanced in future versions, including the introduction of the influence maximization model, the localization of information flow, and the heterogeneity of social groups, among other factors.

- Finally, computational performance is also a bottleneck in the current model. With the increase of the simulation scale, the simulation speed and accuracy problems gradually appear, especially when simulating on large-scale platforms. In the future, optimization algorithms or parallel computing can be considered to improve the simulation efficiency so that it can handle larger data sets and complex interaction scenarios.

- **Future Applications**:
  Our findings have implications for platform governance, AI safety research, and policy-making. The platform can use this simulation to optimize the advertisement detection mechanism to prevent the proliferation of false advertisements and safeguard user experience and platform health. In terms of AI security, this simulation provides a reference for the development of a more robust AI monitoring system to ensure the safety and ethics of platform content. AI-driven moderation strategies must be continuously refined to counteract evolving bot tactics, ensuring that automated systems remain effective in safeguarding digital ecosystems (Bradshaw et al., 2021). As automated moderation continues to shape online discourse, interdisciplinary collaborations between technologists, policymakers, and ethicists will be essential in navigating the challenges posed by AI-driven interactions.

## §6. References


- *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*

- *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 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.