Battle of the AIs: The Arms Race of Generative Content and Detection

Names: Leonel DE AMORIM, Christina PIANG SANG, Vidhi PANDYA
GitHub Link: https://github.com/vidhipandya29/battle-of-the-ais

1. Phenomenon Overview

Our phenomena of interest is the spread and detection of AI-generated deep fake content on social media platforms like Instagram and Meta, with a particular focus on labelling such content and its impact on users behaviour. We aim to investigate if labelling AI generated deep fake images affects its spread and whether labelling affects user behaviour and interactions. The AI-to-AI interactions in this phenomenon is the battle between the AI models to detect AI generated content and the AI bots creating such content and posting them. This AI-to-AI interaction mirrors that of a cat and mouse dynamic in which both the creators and detectors continuously evolve to outpace the other. Major social media platforms such as Meta and Instagram have taken the initiative to label these AI-generated images and videos for transparency. (Meta, 2024) This indicates that this is a present issue and one that has such large platforms working on tackling.

Problem Statement:

The rise of AI-generated deep fake content represents a significant threat to public trust on social media platforms. According to DeepMedia, by the end of 2025, it is expected that 8 million deepfake images and videos will be shared on social media platforms, doubling every six months. (Jacobson, 2024). Although AI generation is growing fast, it can also be used to mislead people on social media platforms. As a result this may lead to growing disinformation narratives and an increase in misleading content. This causes distorted views on social media platforms and is something that must be regulated. Just think that without regulation users on online platforms can not tell what is real content and what is AI generated causing real content and media to be lost among all the AI generated content. Our simulation was created to visualize the relationship between AI generated content and AI-detected content, exploring how user interactions on labeled/unlabeled content impact the spread of this content.

Agent-Based Modelling Approach:

Agent-based modeling is a suitable approach for this phenomenon because it has the ability to model complex and dynamic interactions between designated agents. Given our phenomena there are various interactions at play such as users interacting with the AI content, the AI detection on these platforms and the bot creation of such content. These various interactions between users, generated content and content detection require the use of Agent-Based Modelling to provide a more detailed understanding of how humans interact with content on social media platforms and how such content spreads. In this case, Virus on a Network will be used to show the spread of content that is AI-generated, interactions of users with this content, and how the interactions potentially decrease when it is labelled as AI-generated. Just like how a virus spreads through a network we can think of the AI generated content as a virus and how we can detect and label the virus so it does not spread to users and so users

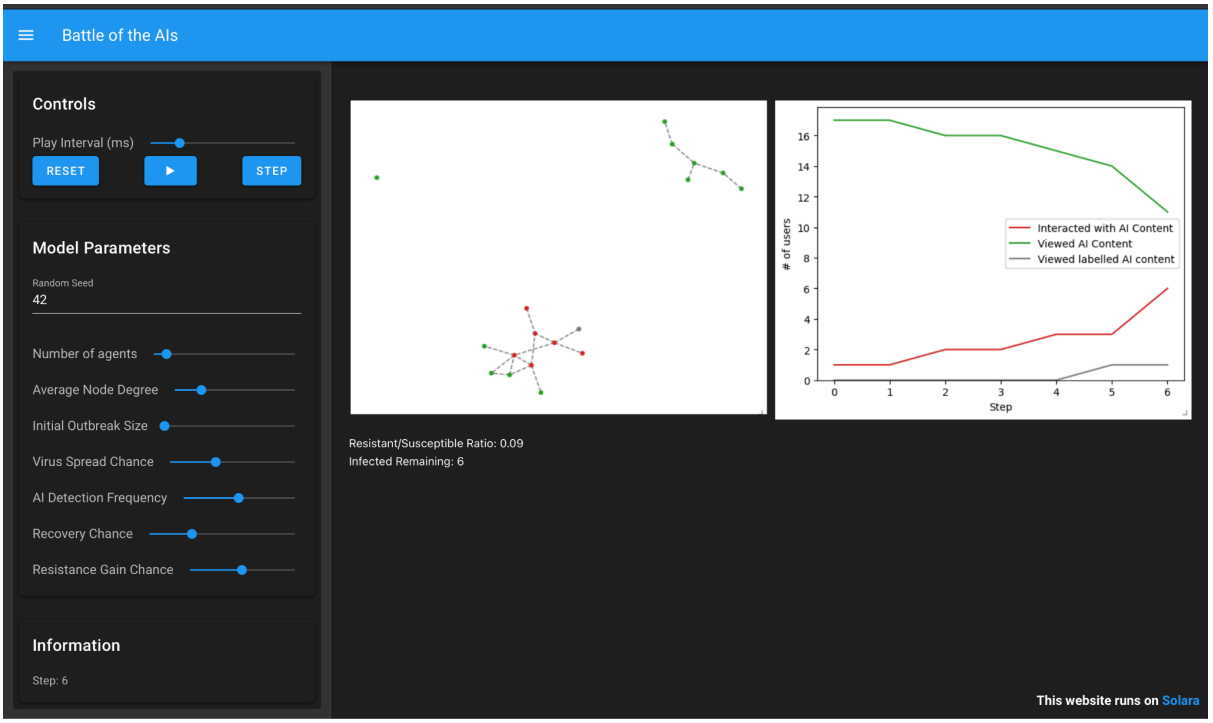
know what they are interacting with. As you can see the Agent-based model of virus on a network provides a powerful model in which we can explore emergent properties and the complex interactions.

Phenomena Illustration

Our agent-based model of AI content spread provides a clear visualization of how deepfakes propagate through social networks and how detection systems respond. The simulation reveals several key patterns that mirror real-world dynamics:

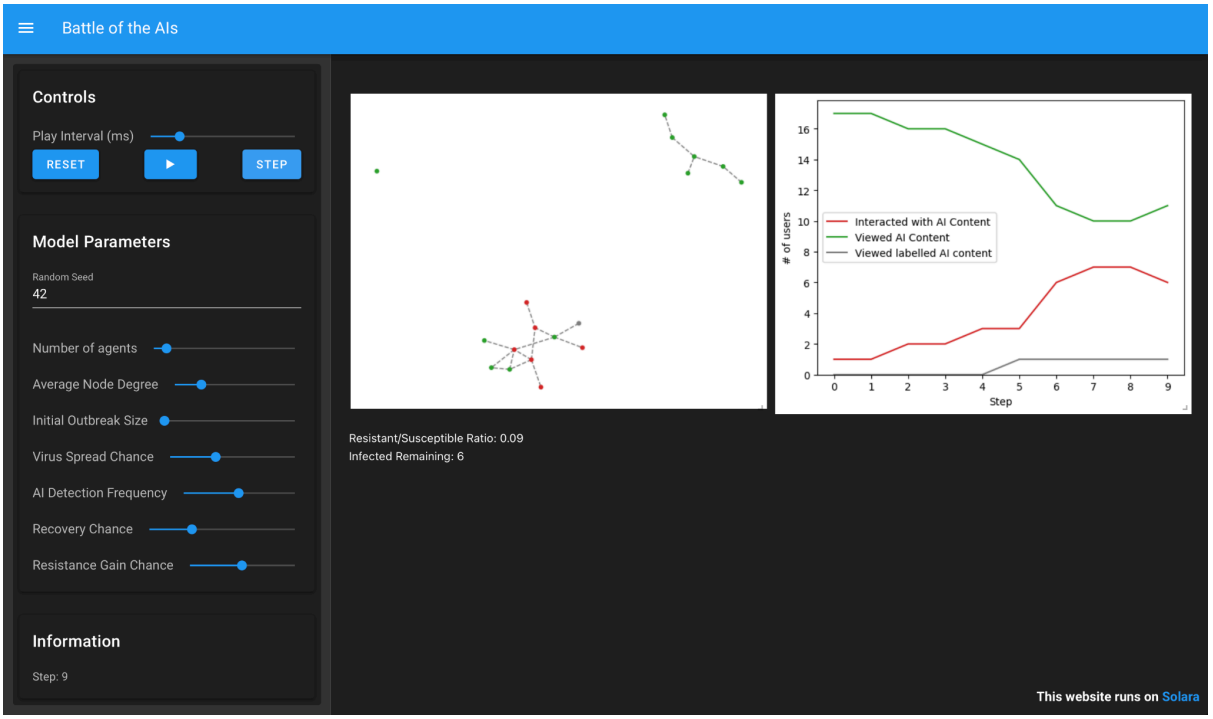
In the initial state, the network begins with primarily exposed users who have not interacted with the content (green nodes) and a small number of users exposed to AI content and who have engaged with content (red nodes). This represents the moment when new deepfake content first enters a social network, similar to when a fabricated political video or celebrity deepfake is first uploaded to platforms like Twitter or TikTok (Vosoughi, Roy, & Aral, 2018). Such dynamics are reminiscent of the earliest stages of a Susceptible–Infected–Recovered (SIR) model, where a small infected seed can initiate a large-scale outbreak (Kermack & McKendrick, 1927).

At Step 6, we see a central cluster of six Infected (red) agents and at least one agent that has become Resistant (gray). Several Susceptible (green) agents remain in other parts of the network, either because they have not yet come into contact with the infected cluster or because random chance spared them from infection. On the right-hand plot, the red line is now climbing to about six infected users, while the gray line (resistant) has begun to appear but remains relatively low. Meanwhile, the green line (susceptible users) is steadily declining from its initial high value. Overall, the infection wave is picking up pace in the center cluster, but the emergence of gray nodes indicates that some users have recovered and become resistant, slowing the outbreak's momentum in those local areas. Similar clustering behavior—where local connectivity accelerates spread while recovery or resistance in certain pockets slows it—has been observed in network-based epidemiological studies (Pastor-Satorras & Vespignani, 2001).

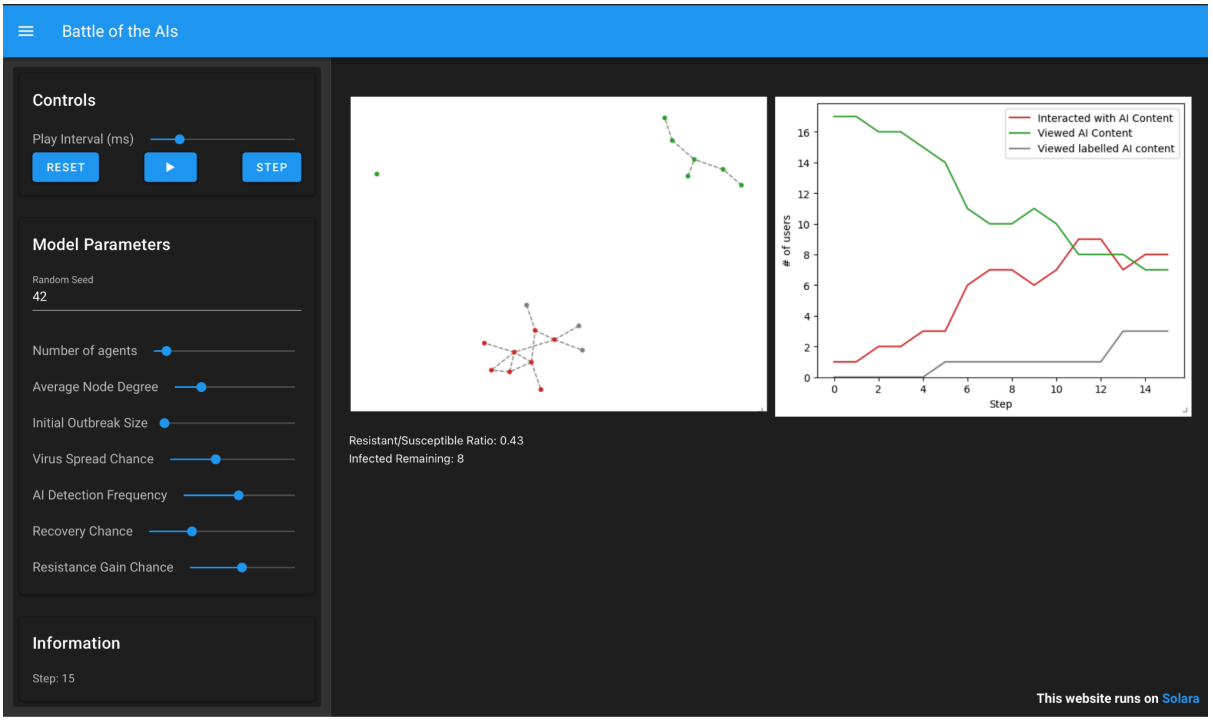


By Step 9, most of the simulation's action has settled around the lower-central cluster, where several agents remain Infected (red) alongside a small handful of Resistant (gray). This local grouping continues to spread the infection internally, especially among any nearby Susceptible (green) nodes that haven't yet been exposed. Meanwhile, the upper section of the network remains almost entirely green, suggesting that the outbreak never migrated into that region—likely due to weak or nonexistent network links.

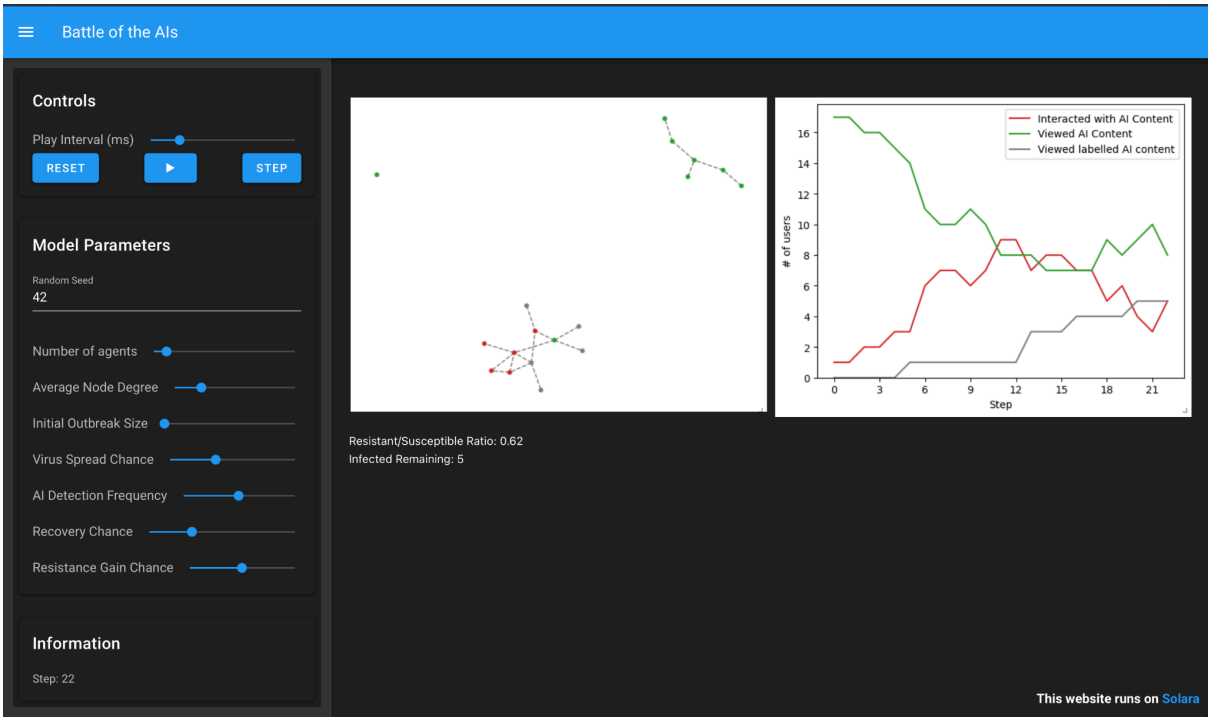
On the right-hand chart, the green line (representing Susceptible users) has steadily declined from its initial high but still hovers around 10 or 11, showing that while many agents were infected over time, a significant number were either out of reach or avoided infection. The red line (Infected) has risen noticeably compared to the early steps but has recently leveled off or dipped slightly, indicating a slowing of the outbreak. The gray line (Resistant) is low, demonstrating that only a few infected users have fully recovered and gained resistance so far. Overall, the infection continues to linger in the lower cluster with six remaining red nodes, but persistent pockets of green in other regions of the network remain largely untouched by the outbreak.



By Step 15, the simulation has settled into a pattern where a core cluster at the bottom of the screen still contains both Infected (red) and Resistant (gray) agents, while the upper portion of the network remains largely Susceptible (green) and isolated from the outbreak. As shown on the right-hand chart, the green line (representing users who have only "Viewed AI Content") has dropped substantially from its early high but has stabilized in the midrange. Meanwhile, the red line (representing users who have "Interacted" or become Infected) has risen to around eight agents and is gradually leveling off; at this stage, the infection no longer spreads explosively because many previously infected agents have turned gray (Resistant), and unconnected green clusters remain out of reach.



By Step 22, the simulation has largely stabilized around a single mixed cluster in the lower portion, where five Infected (red) agents remain alongside multiple Resistant (gray) neighbors. The top clusters, still predominantly Susceptible (green), have been spared from the outbreak due to weak or nonexistent network connections to the infected group. The Resistant/Susceptible Ratio (0.62) indicates there are still more green (Susceptible) nodes overall than gray (Resistant) ones, despite an ongoing wave of recoveries in the main outbreak cluster.



Our simulated patterns of spread, detection lag, and resistance formation align with documented behaviors in social media ecosystems, providing valuable insights into potential intervention strategies for real-world platforms.

2. Simulation Design & Implementation

System Overview

The simulation implementation visualizes the spread of AI-generated deep fake content on a social media platform. It features three key types of agents: regular users, AI-generation bots, and AI-detection bots. The concept of the simulation is a “battle” between two AI systems, generators that create and spread fake content, and detectors that detect and label content. This creates an ecosystem where both human-to-AI and AI to AI interactions influence the growth and reach of content on the social media platform.

Simulation Environment

Our model simulates AI content spread within a network-based environment that represents a social media ecosystem. We implemented this using Mesa's Network class, which provides a flexible framework for modeling complex social interactions. The environment is structured as an Erdős–Rényi random graph, where nodes represent individual users and edges represent social connections between them. This network topology captures the essential characteristics of social media platforms where content can spread from user to user based on their connection patterns. The random graph model creates a realistic distribution of connections, with some users having many connections (influencers) and others having fewer (casual users).Within this environment, information flows along network edges, simulating how content propagates through social relationships. Each connection represents a potential

pathway for AI content to spread, and the structure of these connections significantly influences spread patterns.

Key Parameters:

1. **Number of Users** - The number of users on the social media platform
2. **Average Node Degree** - The number of connections between users displaying a network for users to spread content on the social media platform.
3. **Initial Outbreak Size** - The number of initial users that are infected
4. **Virus Spread Chance -** The chance of users interacting with AI-generated content
5. **AI Detection Frequency -** The frequency of AI being detected
6. **Recovery Chance -** The chance of a user becoming resistant to engaging with AI-generated content
7. **Resistance Gain Chance -** The chance of a user learning to recognize AI-generated context

Agent Design

There are three key agents implemented in the simulation. There are human users on the social media platform, AI content generators that post deepfakes on the platform and AI detectors that label the AI-generated content.

- **Human Users:** These are the users engaging with the content by liking, commenting and sharing..
- **AI Content Generators**: These are the bots responsible for creating and spreading AI Generated deep fake content.
- **AI Detectors:** These are the bots that detect and label AI generated content. They are the adjudicators of deep fake content.

Interaction Dynamics

The simulation uses Mesa's RandomActivation scheduler, which activates agents in random order during each step, creating unpredictable interaction patterns. The model focuses on user interactions with the AI generated content and is presented in the following states:

1. **Susceptible:** Users who have been exposed to AI-generated content but have not interacted with it.
2. **Infected:** Users who have interacted with AI-generated content, contributing to its spread.
3. **Resistant:** Users who have become immune by recognizing AI detection labels, reducing the likelihood of further interaction.

The model tracks the number of users in each state and the changes from resistant to susceptible and infected. The simulation continues to run until a maximum amount of steps is taken or until no other interactions occur. The design is to highlight how content labelling and user behavior are intertwined resulting in a connected influence into the spread of the AI generated content on social media platforms.

Data Collection & Visualization
The simulation collects several key metrics through Mesa's DataCollector:

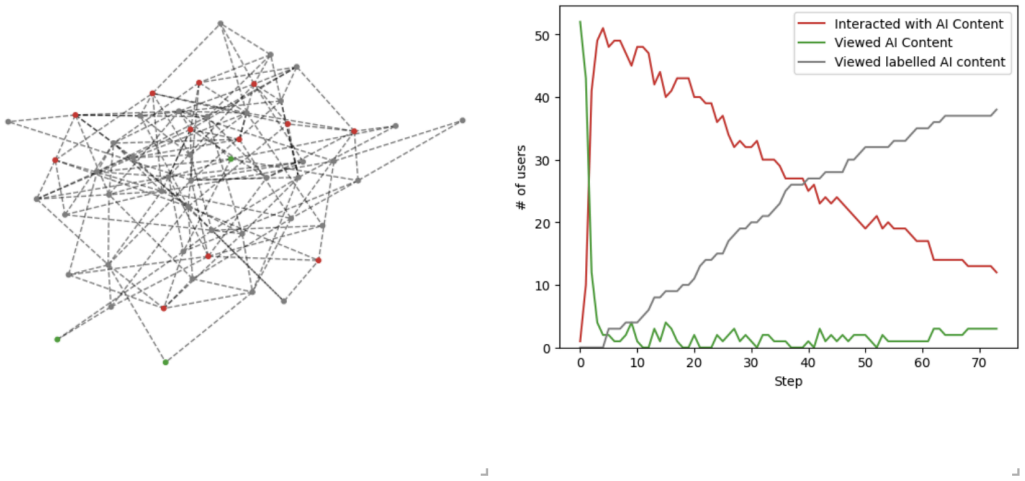
1. **User States**: Tracking Interacted(infected), Viewed(susceptible), and Labeled Content (resistant)
2. **Resistant to Susceptible Ratio**: Monitoring the ratio between the resistant and susceptible users to analyze the effectiveness of content labeling.

Visualization includes:

- Network visualization with color-coded nodes (red for users interacted with AI content, green is for users who viewed AI content but not engaged with it, grey is for users users who viewed labeled AI content and are resistant)
- Time-series chart tracking user states over time showing how content spread and resistance is developed
- Resistance/Susceptible/Infected Ratio that shows the ratio of each state in time as the simulation progresses and how each influences the other.

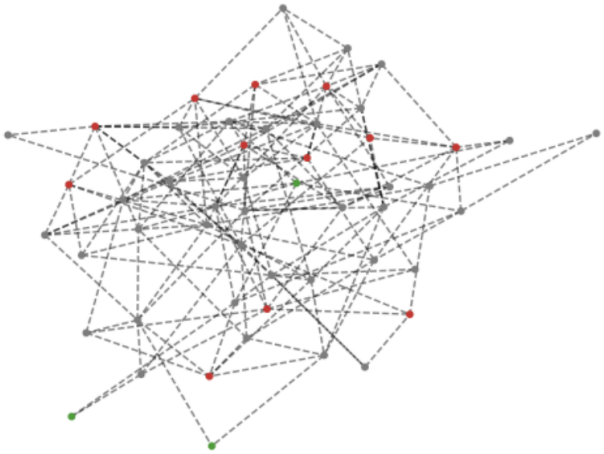
The model collected data reveals that given the labelling there is a resistance that is formed throughout the simulation. Through labelling the spread of the AI content is stagnated and the label allows the users to gain resistance and not interact with the content presented moving from susceptible to resistant.

3. Observations & Results
Simulation Results

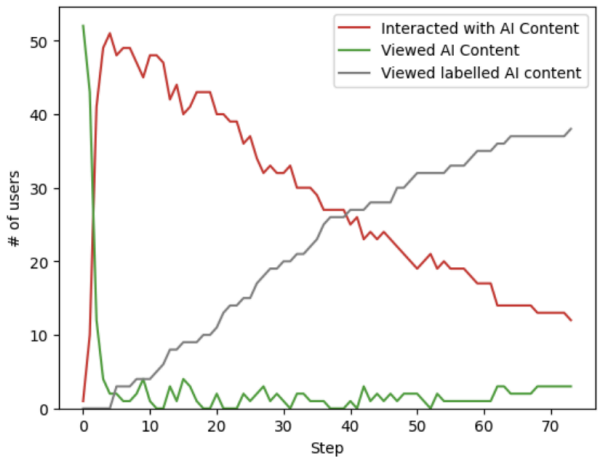


Resistant/Susceptible Ratio: 12.67
Infected Remaining: 12

The visualization for our phenomenon has two key components, the Erdős–Rényi random graph and a time-series graph.



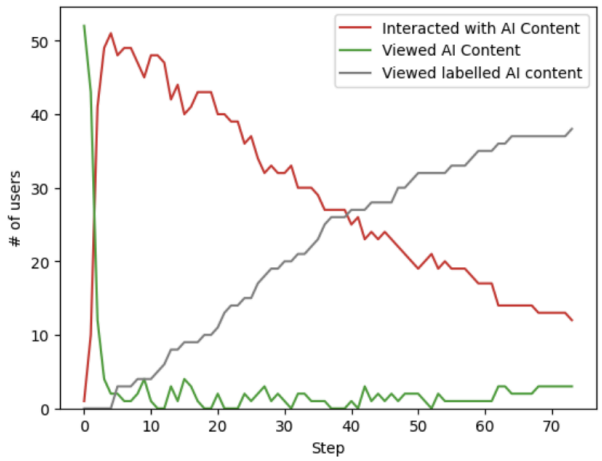
The Erdős–Rényi random graph displays the network on social media platforms. Each node is a user on the platform and the edges display the connections between the users. When users interact with content and change states, this can impact neighbouring nodes spreading the "virus" of AI-generated content.



The time-series graph displays the correlation between the behaviours of the three user agents. The green line displays users that are susceptible to AI-generated content, meaning they have viewed/been exposed to it. The red line displays users that have interacted with AI-generated content. The gray line displays users that are "immune" to the AI-generated content because they are able to see the label.

The simulation also has incorporated how user interactions and social media algorithms work in the real-world. Even if users become resistant, if other users in their network are still engaging with the AI-generated content, they can become infected again to also interact with it.

This can be seen in this graph:



Although users decline in being infected, there is resistance in this decline because of the network dynamics. This is accurate to real-world scenarios because even if you are initially skeptical of a post due to an AI-label, if enough people in your social network interact with it, you will also engage.

Unexpected Behaviours:
It was surprising to see how even though there is friction in adding resistance to users, it still does steadily increase making users susceptible to AI-generated content and not interacting with it.

- This is likely due to the nature of the Virus on a Network framework and how it shows the spread of a virus and its increase in immunity as time passes.

The behaviour of users may be different in real-life situations because it is more unexpected. This is an ideal scenario where if deepfakes are created to spread misinformation, users become resistant to interacting with it over time.

However, in real life, individual perspectives, trust and social media algorithms play a key role in interactions with content.

4. Ethical & Societal Reflections

- **Ethical Considerations:**

This simulation has brought out some very important ethical considerations regarding the use and spread of AI-generated content. Although we did not use real-world data in this model, it did not have any personal user information, social media interactions, or content. The agent-based simulation is still based on the same dynamics that mimic actual patterns of AI media diffusion, as it mirrors ethical risks that could arise from deepfake content creation and dissemination on such a large scale.

An ethical concern central to this includes the usage of personal images without any knowledge or consent. Deepfake contents usually use AI models fed with images, videos, or voice data of people that are publicly available, and most of these are gathered from platforms like Instagram, TikTok, or YouTube. This is a colossal threat to digital autonomy, as people might not even be aware of the fact that their own likenesses are being made to perform whatever the content demands. Although our model is abstract, it does show the mechanics that make it possible. A notable real-world example includes the "Zao" app controversy in 2019, where users unknowingly consented to the broad reuse of their images,

highlighting significant ethical violations related to informed consent, digital autonomy, and privacy (Heaven, 2019).

This connects directly to privacy concerns and ethical principles such as informed consent, dignity, and the right to be forgotten, as outlined in frameworks like the *Montreal Declaration for Responsible AI* (Université de Montréal, 2018) and *OECD AI Principles (OECD, 2019)*. These frameworks emphasize transparency, accountability, and respect for fundamental human rights, which are all relevant when modeling the spread of manipulated media.

Our simulation, which analyses the behavioural impacts of deepfake content particularly how people engage with such contents as well as how they resist them has useful implications for the role of a seemingly neutral AI system in undermining trust, amplifying harm and enabling exploitation if not framed with ethical safeguards.

- **Societal Implications:**

Micro-Level: At the individual level, our simulation demonstrates how AI-generated deepfake content can harm human identity and reputation. Individuals may be victims of identity theft, defamation, or malicious impersonation. The emotional discomfort and sense of powerlessness that come with having one's likeness modified without consent have serious psychological and social effects. For example, high-profile events using deepfake impersonations of political figures and celebrities highlight the potentially harmful personal consequences of these technologies (Chesney & Citron, 2019).

Meso-Level: On a broader societal or community level, the spread of AI-generated misinformation contributes significantly to a decline in public trust within digital platforms and media ecosystems. Our simulation illustrates how rapidly misinformation can proliferate, reflecting real-world scenarios such as coordinated disinformation campaigns observed during elections or public health crises, including the COVID-19 pandemic (Allcott et al., 2019). Effective labeling significantly mitigates this risk, as our simulation demonstrated, emphasizing the critical role of transparency and user awareness in maintaining societal trust.

Macro-Level: At the structural and policy levels, our simulation findings reinforce the necessity for strong regulatory frameworks and platform accountability. Effective labeling and content moderation procedures emulated in our study are directly related to real-world governance initiatives such as the EU's AI Act (European Commission, 2021), which requires openness and accountability in the use of AI-generated information. Similarly, social media platforms like Meta, Google, and X (previously Twitter) are increasing their investment in advanced detection and tagging systems to combat misinformation. Our findings highlight the importance of proactive governance and the need for strict regulatory standards to monitor and minimize the societal risks posed by AI-generated content.

Alignment with real-world patterns:

Our model also aligns with real world patterns as well. During early stages of the content being first shared we see a drastic increase in the amount of users interacting with it. This mirrors the viral nature of online posts on social media platforms like Facebook, Instagram and TikTok. Images and AI generated content will drive user engagement when it is viral allowing more users to be exposed to creating a cycle spreading the content. We also see how when effective labelling is applied that the spread of such content is drastically reduced as users become aware of the content they are interacting with and ultimately decide to interact with the content less.Labelling results in a form of resistance as users become less likely to engage with AI deep fake content as they are exposed to the labels. This is consistent with empirical studies showing reduced user interaction with labeled misinformation (Pennycook et al., 2020).

While our simulation may not capture all the complex and intricate factors in the real world it does provide a general view of the impact of labelling AI generated content and the impact that has on the individual users actions.

Potential for malicious use:

While we used our simulation to understand how AI content is spread and the impact detection has, it could potentially be repurposed for malicious intents as well.This potential misuse underscores the importance of adhering to ethical AI development principles, such as those detailed in the Asilomar AI Principles (Future of Life Institute, 2017), which call for responsible AI innovation, proactive impact assessments, and stringent oversight mechanisms.

For example by using our model a bad actor could look into the detection lag in certain social media platforms. The model can be used to see how long it takes for a platform's AI detection software to detect the AI generated content, this would allow them to identify how long the system takes to detect and they can then design bots to post such content and delete them before detection is raised. This would allow bad actors to post constant media content while never being detected, allowing for the malicious content to be spread reaching users all while not being detected by the system. It could also be used to test what content is being detected by the AI system allowing bad actors to identify what such AI detection systems are looking for and allowing them to evade detection. So while our model is used to illustrate the impact of labelling AI content on social media platforms it can also be used to aid those who aim in spreading AI content.

5. Lessons Learned & Future Directions

Design and Development Reflections

There were some challenges we faced while designing and implementing our agent based model for our phenomena. Initially while developing our model we struggled with the visualization of our simulation, we wanted to implement a system that would accurately represent the complex dynamics of a social media ecosystem. We debated between the Erdős–Rényi random graph and the Watt Strogatz Graph. We ultimately decided to implement the Watt Strogatz Graph as we found that it would allow us to visualize clustering which would be useful in representing close knit groups on social media platforms. This allowed us to visualize the interconnections among clusters and random users as well. We also faced the challenge of adapting the MESA model to fit our project needs. While the MESA models provide a great foundation to work with, it still required a considerable amount of customization in order to find the dynamics we wished to present. We modified the model to be able to illustrate the AI content bots, user interaction and the label and detection implementation. This proved challenging as we had to constantly tweak the implementation to be able to accurately represent the social interactions we needed to illustrate in our model. Despite these challenges we as a team managed to work collaboratively to address them. We regularly discussed what our desired outcome was and how we could make adjustments to reach this outcome. We focused on testing our model to ensure it closely aligned with our goals and ensuring our model was able to illustrate the spread of deepfake content and the impact labelling such content had on the content spread and user behaviour on social media platforms.

Model Limitations & Areas for Improvement

While our agent based model provides a visualization into the dynamic interactions between user, AI detection systems and AI content creation bots we feel it still has some limitations and areas for improvements. For example, a significant constraint is the oversimplification of social media interactions. In reality there is a complex dynamic of interactions involved on social media platforms and the users on them who engage with the content. Factors such as demographics, trends, news cycles, and platform engagement algorithms are all key factors that impact the way users interact with content and the manner in which content spreads on a platform. Our model simplifies these interactions given we do not have the data present to create a model to exemplify these interactions, we believe that this is an area of improvement as a more complex and intricate model accounting for these and other factors would allows for a far more in depth visualization of content on a social media platform, the users engagement and the impact labelling would have. Another limitation is not being able to implement a model that takes into account social interactions such as liking, commentating and sharing. Each would have a different value in terms of its influence on social interactions. These types of interactions would have been great to take into account to visualizes social aspects such as, perceived authenticity given a post comes from a friend or family member, and social pressures of engaging with content deemed viral.These are all interaction that we be great to take into account to understand what impacts a users willingness to interact with

AI generated content and contrast that with content that is labelled. We also illustrated a broad implementation of AI detection to catch AI generated content on a platform and an improvement would have been to incorporate real AI detection models that platforms use to truly understand real world detection models and how their detections work and the correlation they have to the spread of AI generated content. A more complex detection model would allow us a deeper, more accurate representation of social media ecosystems and how they can be improved to allow for accurate and effective content labelling.

Future Applications

Given that through our simulation we were able to see that labelling does in fact impact users willingness to interact with AI generated content we believe these findings can be used in research fields to identify and understand how AI generated content and detection systems interact with human users. What are the key factors that impact user interactions? And how can these factors be targeted and used? This could lead to a development of better AI detection guidelines and of AI ethics on social media platforms. This can be used to ensure that AI technologies are deployed to allow for greater public transparency and public trust on social platforms. We can use this simulation to create future regulations and guidelines regarding the use of AI generated content on social media platforms. By understanding how AI generated content spreads and the influence they can have on public perception then we can create an effective framework for social media platforms to follow in regard to content moderation and disinformation control. Enforcing social media platforms to label AI content through the use of policies and legislation would allow for users to be protected against misinformation campaigns and allow for an increase in social media transparency.

References

1. *Labeling AI-generated images on Facebook, Instagram and threads*. Meta. (2024, February 14). https://about.fb.com/news/2024/02/labeling-ai-generated-images-on-facebook-instagram-and-threads/
2. Jacobson, N. (2024, February 26). *Deepfakes and their impact on society*. CPI OpenFox. https://www.openfox.com/deepfakes-and-their-impact-on-society/#:~:text=According%20to%20DeepMedia%2C%20in%202023,doubling%20deepfakes%20every%20six%20months
3. *Our approach to labeling AI-generated content and Manipulated Media*. Meta. (2024, September 12). https://about.fb.com/news/2024/04/metas-approach-to-labeling-ai-generated-content-and-manipulated-media/
4. Farrell, D. (n.d.). *Agent-based models in Python with Mesa and NetworkX.* Retrieved from https://dmnfarrell.github.io/bioinformatics/abm-mesa-network
5. Mesa Team. (n.d.). Virus on a Network. Mesa Documentation. Retrieved [Date], from https://mesa.readthedocs.io/latest/examples/basic/virus_on_network.html
6. Readthedocs. (n.d.). https://mesa.readthedocs.io/_/downloads/en/stable/pdf/
7. *Matplotlib.axes.axes.get_legend_handles_labels#.* matplotlib.axes.Axes.get_legend_handles_labels - Matplotlib 3.10.1 documentation. (n.d.). https://matplotlib.org/stable/api/_as_gen/matplotlib.axes.Axes.get_legend_handles_labels.html
8. Luna, E. de. (2024, June 24). *Photographers say meta is labeling their photos as being "made with ai" even when they're not*. Mashable. https://mashable.com/article/meta-made-with-ai-label
9. Dogra, R. (2024, December 26). *AI-generated content surges on social media: New Study reveals Startling Trends*. AI World Today. https://www.aiworldtoday.net/p/research-shows-ai-generated-content-surges-on-social-media
10. Miles, M. (2025, January 24). *How to write a problem statement (with 3 examples)*. BetterUp. https://www.betterup.com/blog/problem-statement
11. de Seta, G. (2022). *Huanlian, or changing faces: Deepfakes on Chinese digital media platforms*. New Media & Society, 27(4), 739–756. https://doi.org/10.1177/13548565211030185
12. Université de Montréal. (2018). Montreal Declaration for Responsible AI.
13. OECD. (2019). OECD Principles on Artificial Intelligence.
14. Chesney, R., & Citron, D. (2019). Deepfakes and the New Disinformation War: The Coming Age of Post-Truth Geopolitics. Foreign Affairs.
15. Allcott, H., Gentzkow, M., & Yu, C. (2019). Trends in the diffusion of misinformation on social media. Research & Politics.
16. European Commission. (2021). Proposal for a Regulation laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act).

17. Future of Life Institute. (2017). Asilomar AI Principles.
18. Pennycook, G., Rand, D. G., & Bak-Coleman, J. (2020). Fighting misinformation on social media using crowdsourced judgments of news source quality. Proceedings of the National Academy of Sciences.
19. Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E., & Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, 113(3), 554–559.
20. Kermack, W. O., & McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society A*, 115(772), 700–721.
21. Pastor-Satorras, R., & Vespignani, A. (2001). Epidemic spreading in scale-free networks. *Physical Review Letters*, 86(14), 3200–3203.

Attestation

Name	Contributions
Christina Piang Sang	-Reviewed Sections for improvements and additions -Implemented Section 4,5,6
Leonel De Amorin	-Reviewed Sections for improvements and additions -Implemented Section 4,5,6
Vidhi Pandya	-Reviewed Sections for improvements and additions -Implemented Sections 1, 3 - Updated code