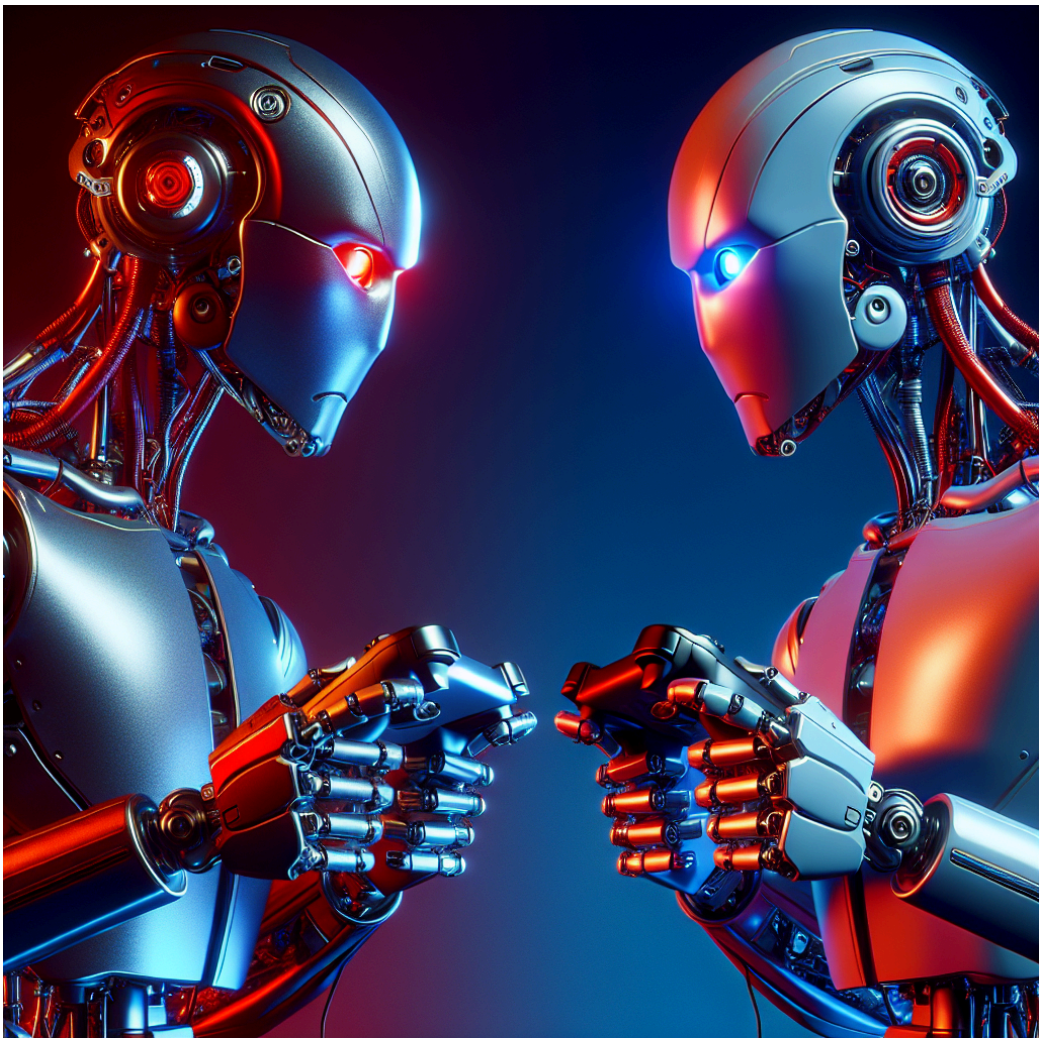
AI Undercover

AI Bots in the world of AI Cheat Detection



Faraz Akbarzadeh (*Leader/Facilitator*)
Adil Guluzade (*Checker/Editor*)
Santusht Arora (*Devil's advocate*)
GitHub: https://github.com/adilgulu/EECS4461

Phenomenon Overview

The media environment undergoes rapid transformation through artificial intelligence which reshapes digital interactions at their core. The continuous battle between AI-enabled cheating systems and their anti-cheat detection counterparts represents a significant and growing trend in media transformation. The online gaming environment showcases a complex interaction where cheating AI updates its methods to avoid detection and anti-cheating mechanisms develop new strategies to recognize and neutralize these advanced threats (Chen, 2024; Skinner & Walmsley, 2019).

Our simulation focused on this phenomenon by using a scenario derived from the online multiplayer game Slither.io. Players maneuver snake avatars in Slither.io which expand their length through pellet consumption in order to achieve the longest snake length. Cheating AI bots take advantage of the game mechanics to speed up their growth artificially which gives them an unfair edge by growing faster and staying in the game longer. AI-powered cheat detectors work by examining growth patterns to detect and mark abnormal behaviors that suggest cheating. The particular environment described here establishes a practical basis for examining AI-to-AI interaction patterns within media systems.

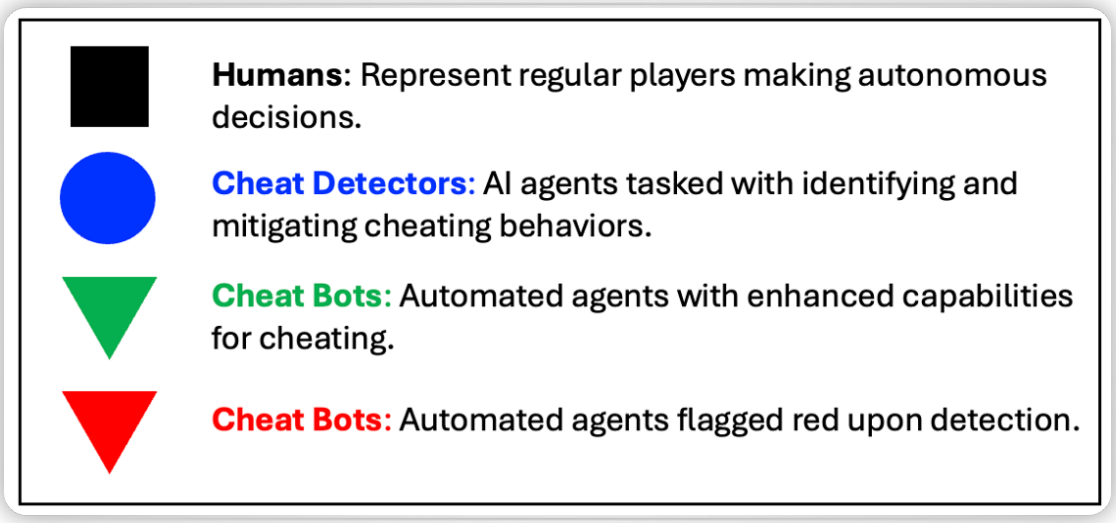
The primary importance of AI-to-AI interactions stems from their substantial influence on digital trustworthiness and the equitable function of online platforms. Cheaters who continuously break the rules damage fair play while simultaneously destroying user confidence which leads to lower engagement and threatens digital economy sustainability. Anti-cheat systems that are too strict or badly constructed pose a risk of punishing real players inaccurately which leads to user dissatisfaction and abandonment thereby harming both platform engagement and community well-being (Lehtonen, Vesa, & Harviainen, 2022).

The rapidly increasing complexity of AI-driven cheating mechanisms makes it essential to maintain a delicate balance between effective cheat detection and user fairness. Cheating agents utilize complex methods like exploiting game mechanic flaws and advanced vision-based cheating techniques to gain unnoticed advantages beyond traditional detection capabilities. The use of adaptive methods requires ongoing improvements to detection strategies to advance anti-cheating technology toward stronger and smarter proactive solutions according to Jonnalagadda et al. (2021).

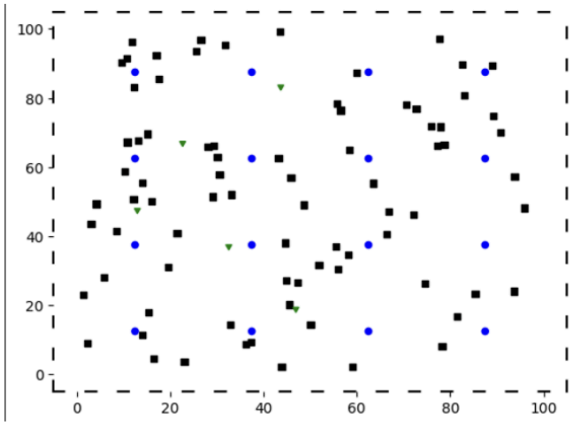
Agent-based modeling (ABM) stands out as an effective research methodology because it enables the simulation of intricate interactions between autonomous agents. The nuanced behavioral dynamics captured by ABM in defined rulesets make it particularly suitable for analyzing adaptive behaviors found in AI cheat and anti-cheat interactions. Through explicit modeling of individual agent behaviors which include human users (black), cheat detectors (blue), cheating bots (green), and flagged cheaters (red)—we can analyze how these small-scale interactions build up to create large-scale emergent effects.

Our study of this phenomenon is best demonstrated through sequential annotated visualizations that show simulation phases in a comic-style layout.

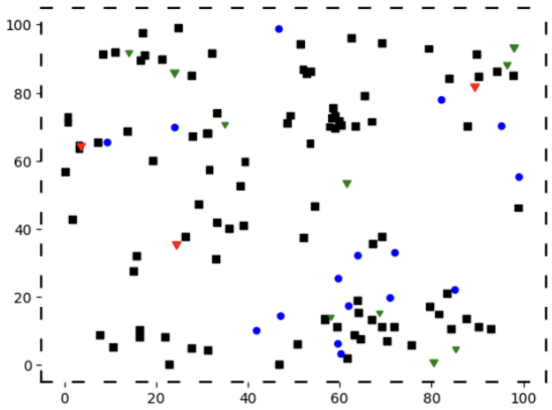
- **Step 0:** Initially, agents are randomly distributed with no active detection. There are 79 regular players, 16 cheat detectors, and 5 cheaters. No agents have been flagged or eliminated yet.
- **Step 40:** Detection becomes active. Cheat detectors identify and flag cheaters, with 4 flagged cheaters emerging. At this stage, there are 87 regular players, 16 cheat detectors, and 12 cheaters. Cheat bots reach a high score of 21.1, compared to 6.08 for regular players. True positives stand at 4, with 8 false negatives.
- **Step 120:** Detection has advanced significantly. Regular players number 111, cheat detectors remain at 16, and cheaters reduce to 8, with only 1 flagged. Cheat bots achieve a high score of 22.28, whereas regular players achieve 17.78. Detection efficacy remains strong with 1 true positive and 7 false negatives.
- **Step 248:** The detection system reaches maturity. Eliminations include 118 regular players and 35 cheaters. Remaining agents are 92 regular players, 16 cheat detectors, and only 4 cheaters with no flagged cheaters. Regular player scores peak at 40.21 compared to 18.27 for cheat bots. Detection shows slight inaccuracies with 2 false positives and 4 false negatives.



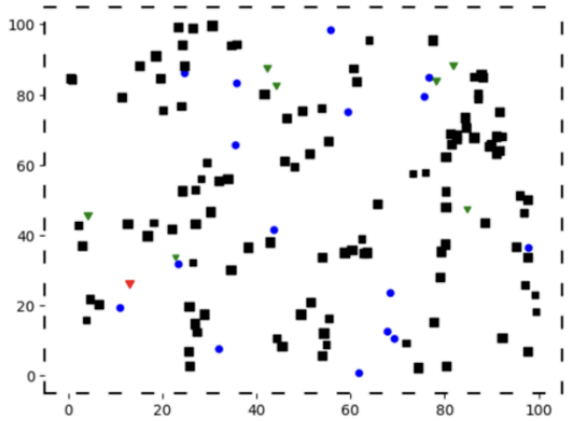
Step 0



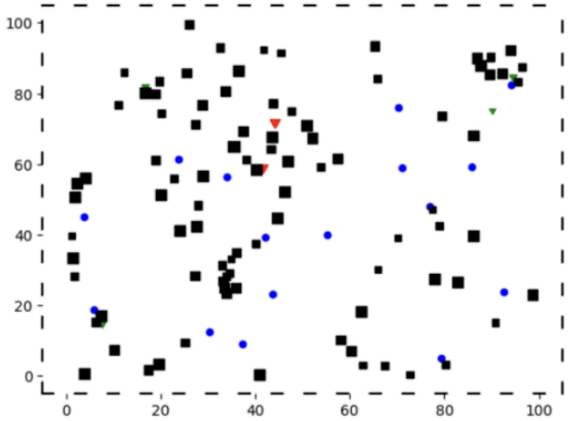
Step 40



Step 120



Step 248



Simulation Design & Implementation

System Overview: This project simulation models complex interactions between three main agent types which include standard players resembling humans, cheat bots operated by AI technology, and cheat detection bots also controlled by AI systems. The simulation aims to investigate how AI-to-AI interactions produce emergent behavior patterns in a competitive game environment based on Slither.io.

This simulation features regular human-like agents which are displayed as black squares and these agents mimic real human players by moving randomly throughout the environment while collecting scores through conventional gameplay actions. AI cheat bots which appear as green triangles use intentional game mechanic exploits to achieve score growth rates that surpass those of human players. The blue circles represent cheat detection bots which focus

solely on monitoring players for unusual growth patterns that signal cheating activities. The cheat bots receive a red flag when they are recognized as compromised. The perpetual process of detection and evasion followed by adaptation builds an evolving cycle resulting in complex emergent behaviors that define AI-to-AI strategic contests.

Simulation Environment: The simulation environment functions as a continuous toroidal space made possible through the ContinuousSpace module from Mesa. The design choice creates a virtual environment without boundaries similar to open-world multiplayer games where player boundaries are non-existent. When agents move beyond one boundary of the simulation they appear on the opposite side which removes edge-related biases and accurately models real-world online digital environments.

Key parameters defined within the environment include:

- **Agent Counts:** At the start of the simulation regular users and cheat bots along with detection bots each have predefined initial populations.
- **Agent Movement Speeds:** The movement speeds are adjusted to show differences in how quickly agents respond and carry out their strategies.
- **Detection Radius:** The detection radius parameter establishes the monitoring range for cheat detection agents when observing other agents.
- **Scoring Thresholds:** We set thresholds to distinguish legitimate scoring patterns from potential cheating activities.

Agent Design: Each category of agents receives careful design to portray unique roles and distinct behaviors.

- **Regular Players (Black Squares):** These agents demonstrate random path movements and standard incremental points which reflect the behavior of typical players in the game. The agents maintain basic operations to create a standard for authentic player behavior in the system.
- **Cheat Bots (Green Triangles):** These agents operate through enhanced growth algorithms that target game mechanics to achieve disproportionate score gains compared to standard players. The scoring benefit these agents gain replicates cheating tactics that have been established in academic studies (Jonnalagadda et al., 2021).
- **Cheat Detectors (Blue Circles):** These agents actively monitor real-time scoring rates of nearby agents. The system identifies cheating patterns by examining score irregularities through proximity-based analysis. The system flags agents proven to be cheating by changing their status to red to visually represent their detection status.

The system simplified complex agent behaviors including advanced decision-making and visual pattern recognition into basic proximity and scoring rate measurements to handle computational demands. The implemented simplifications preserved essential behavioral realism needed for precise simulation results and maintained both computational feasibility and effective performance during simulations.

Interaction Dynamics: Mesa's RandomActivation scheduler guides the interactions between agents during the simulation. The simulation chose this scheduling method because it effectively recreates the unpredictable conditions found in online multiplayer games. The random activation method produces unbiased dynamic interactions that successfully simulate unpredictable encounters between human users and AI-controlled agents.

The model primarily bases interactions on agent proximity and scoring evaluations. Cheat detectors perpetually monitor their environment to evaluate the scoring increase rates of nearby players. An agent that crosses the cheating behavior thresholds identified by a cheat detector receives an immediate flag. Agents that are marked for cheating change their movement patterns to evade detection which results in ongoing dynamic cycles of evasion and detection.

The interaction dynamics rest on two fundamental assumptions: detection agents operate with restricted information access which limits their evaluation to adjacent scoring patterns and they do not have direct communication pathways between one another. The constraints implemented in this simulation accurately represent the operational limitations of real-world anti-cheat systems on digital platforms to produce trustworthy results.

Data Collection & Visualization: A thorough examination of emergent phenomena within the simulation required the establishment of extensive data collection practices. Collected data encompassed: Agent Scores: Continuous monitoring evaluates scoring patterns to distinguish authentic gameplay from potential misuse. Flagging Incidents: The documentation process tracked the specific times and agents that received cheating flags. Detection Metrics: Accurate detection assessments require true positives and false positives as well as true negatives and false negatives which serve to pinpoint performance strengths and weaknesses.

Visualization served as an essential tool for understanding complex data patterns and emergent behaviors. Agent positions and statuses were mapped using scatter plots which clearly demonstrated the spatial dynamics of agent interactions. Time-series graphs displayed dynamic trends of agent behaviors while revealing detection efficacy and system responses throughout different periods.

Real-time data processing and visualization clarity presented computational challenges despite the strategic use of visualization techniques. A substantial amount of work went into refining procedures for real-time data acquisition and updates to maintain both the precision and clarity of visualizations.

The model underwent innovative changes by incorporating adaptive agent behaviors which enabled cheat bots to dynamically modify their evasion tactics according to detection rates. The technical enhancements delivered substantial improvements to simulation realism by

replicating complex real-world behaviors validated by academic research on AI adaptability in gaming contexts (Chen, 2024; Skinner & Walmsley, 2019).

Our simulation model reaches high technical precision and deep analytical capacity through refined design processes and iterative improvements complemented by detailed visualization methods while revealing important data about dynamic AI interactions in digital environments.

Observations & Results

Illustrating the Phenomenon of Interest: The simulation results clearly show how AI-to-AI interactions develop in online gaming through the adaptive interactions between cheating bots and anti-cheat systems. As the simulation advances through its different phases, the fundamental phenomenon of adaptive evasion versus detection becomes apparent. The central observation demonstrates the evolutionary battle between cheat bots that adjust to detection methods and cheat detectors which consistently upgrade their methods to stay effective.

Quantitative Metrics and Emergent Behaviors: We systematically investigated the phenomenon by gathering quantitative data on agent states as well as their scores and detection accuracy together with emergent behaviors. The primary data collected included:

- **Number of Regular Players, Cheat Bots, and Detectors:** By monitoring population fluctuations we gain insights into survival rates and detection effectiveness.
- **Scores and Growth Rates:** This analysis examines the performance differences between cheating bots and regular players and tracks the score patterns of agents that have been flagged.
- **Detection Accuracy:** The detection accuracy measurement incorporates both true positives and true negatives as well as false positives and false negatives.
- **Adaptive Evasion Metrics:** We assess the rate at which flagged cheat bots manage to avoid detection.

The system starts at Step 0 with a composition of 79 regular players alongside 16 cheat detectors and 5 cheat bots. At this moment no detection events have transpired and all detection scores remain at zero. This step creates the foundational distribution for agents and their interaction capabilities. Zero detections of cheaters indicate that detection mechanisms were not yet in place.

The detection phase begins at Step 33 as cheat detectors successfully identify and flag the initial group of cheat bots. The detection method identifies 4 cheat bots from a total of 12 active cheaters. The top score for regular players reached 6.08 during this period while cheat

bots achieved their highest score at 21.1. The detection results include 4 true positives and 8 false negatives which demonstrate the initial challenges faced to achieve accurate detection.

The system advances to a more developed detection phase at Step 120. The system has successfully removed 18 cheat bots but still identifies one remaining flagged bot. Active cheat bots decreased to 8 due to effective detection and removal operations. The difference in peak scores demonstrates this shift since regular players top at 17.78 and cheat bots reach 22.28 despite continuous evasion attempts. The single remaining true positive demonstrates cheat bots have modified their tactics to avoid detection as shown by the existence of 7 false negatives.

The system shows stabilized detection mechanisms at Step 248 which represents the final stage being analyzed. The detection system eliminated 35 cheat bots but 4 remain active because they have evolved sophisticated evasion techniques. Standard players reach maximum scores of 40.21 but surviving cheat bots only manage to achieve top scores of 18.27. Detection accuracy decreased slightly because of 2 false positives and 4 false negatives which demonstrate continuous challenges when trying to recognize adaptive bots.

Interpreting Results and Emergent Dynamics: The key emergent behavior detected is the way cheat bots learn to avoid detection through adaptive evasion techniques. Detection begins successfully with numerous true positives but becomes less effective over time as cheat bots adapt by hiding in user groups or changing their behavior patterns occasionally. The strategic adaptation of cheating systems to bypass detection algorithms mirrors real-world observations about system evolution as documented by Jonnalagadda et al. (2021).

The creation of specific clusters for cheat detection resulted in localized improvements in detection accuracy because more concentrated patrols increased the chances of identifying cheaters. The adaptation created unprotected areas where cheat bots continued to operate undisturbed.

The scoring patterns of adaptive cheat bots produced an unexpected outcome during analysis. When cheat bots were marked but remained active, their scoring speed decreased significantly showing a deliberate strategy to lower detection risk. Researchers did not foresee this adaptive behavior which demonstrated bots implementing complex strategies to endure detection by valuing survival over fast score gains when they become flagged.

Potential Explanations and Implications: The dynamic balance between detection and adaptation explains the changes in accuracy metrics. Cheat bots adjust their behaviors after detection events which results in a brief reduction of detection accuracy. Real-world anti-cheat systems face significant difficulties because they need constant updates to stay effective against evolving threats.

The local benefits of clustered cheat detectors indicate that spreading detection mechanisms across multiple locations could improve overall performance. The strategic balance between

patrol density and distribution serves as an effective method to close the gaps which adaptive bots typically exploit.

Going Above and Beyond: Our study uses sophisticated data analysis methods to bridge quantitative measures and new behavioral patterns. Analyzing true positive rates versus adaptive evasion rates revealed strategic weaknesses in the detection algorithm which prompted recommendations to optimize agent placement and patrol procedures.

Our findings advance knowledge about AI interactions in digital systems while emphasizing the need for adaptable detection methods in online games. Enhancing detection systems through predictive algorithms designed to proactively counter bot evasion methods will maintain their effectiveness against evolving threats.

Ethical & Societal Reflections

Ethical Considerations: A primary ethical consideration for our simulation project centers around how we manage data and protect privacy. Our simulation avoided direct privacy violations because it didn't use actual data sources such as social media interactions or player behavior logs. The model functioned entirely within a synthetic framework built on theoretical principles together with simulated agent behaviors which eradicated any potential risks to personal data misuse or privacy violations.

The ethical concerns present in this situation go beyond the simple use of data. The simulation replicates real gaming environments which use cheat detection algorithms to enforce fair play. The possibility of false-positive detections which incorrectly identify legitimate players as cheaters represents a significant problem. The outcome of real-world situations may result in users becoming dissatisfied with the service while facing account suspension and diminished trust in platform reliability. During our simulation we encountered challenges as false positives were generated occasionally, which highlighted the necessity for stronger validation systems in AI-based detection technologies.

The simulation brings up ethical questions regarding the responsibilities of developers when they create and implement anti-cheat systems. Detection systems must decide between reducing false positives through caution or maintaining game integrity by aggressively identifying suspicious behaviors. The issue represents a wider challenge in AI governance because maintaining both security measures and fairness standards creates ongoing disputes. The simulation demonstrated how cheat bots use adaptive evasion tactics which revealed the ethical dangers when developers implement aggressive detection protocols that mistakenly penalize non-cheaters with unusual legitimate behavior.

Societal Implications: Our study reveals the societal implications through its focus on the interactive dynamics between bots and humans in media ecosystems. Online gaming

environments display immediate effects at the micro-level through their need to maintain a balance between fair play and avoiding false accusations. When cheat detection systems maintain lenient standards they enable cheating to spread and damage user trust. Systems that enforce overly strict measures create an environment where legitimate players feel excluded leading to weaker community relationships and reduced player participation.

The results of our study show important consequences for platform governance at the meso-level especially in digital environments where AI moderation and cheating detection systems play a crucial role. Cheat bots' capacity to modify themselves and sidestep detection proves that fixed or responsive anti-cheat solutions do not work well. Platforms need to implement dynamic detection systems which adapt as new cheating methods develop. Platforms need ongoing system updates and machine learning tools paired with community input channels to enhance detection algorithms while maintaining fairness for users.

Our simulated phenomenon at the macro-level illustrates wider challenges faced in AI governance together with digital trust. The increasing prevalence of AI-driven systems in online interaction moderation results in a significant rise in malicious repurposing risks. Real-world malicious software development could benefit from understanding the adaptive evasion tactics exhibited by cheating bots in our simulation. Research guidelines must be established to prevent the use of simulation outcomes for creating superior cheating mechanisms.

Repurposing Concerns: Our simulation framework could be used for malicious purposes which demands careful consideration. Our main goal to study and prevent cheating methods carries an inherent danger that findings from adaptive cheat bot simulations could be leveraged to improve cheating techniques. Our research follows ethical guidelines focused on responsible AI development to address this risk. The development of cheat detection algorithms requires full transparency documentation along with peer reviews and responsible disclosure procedures to ensure they cannot be misused.

Going Above and Beyond: Our reflection finds its basis in both the technical simulation results as well as an in-depth analysis of ethical standards and AI management guidelines. Our simulation outcomes combined with real-world gaming controversies driven by excessive cheat detection measures demonstrate why AI implementations must maintain balance and transparency. Research going forward must investigate cheat detection technology advancements and the societal and ethical structures that direct their deployment in practical environments.

Lessons Learned & Future Directions

Design and Development Reflections: The development of our agent-based simulation model encountered multiple technical and conceptual obstacles across its entire creation process. Our main difficulty involved embedding realistic behavioral dynamics into the simulation without sacrificing computational efficiency. The need to simulate numerous agents with diverse behaviors resulted in higher computational demands which caused problems including system lag and delayed processing. We improved the model's performance by decreasing non-essential update occurrences while using advanced data structures to manage agent states.

Accurately modeling the adaptive behavior of cheating bots posed a significant conceptual difficulty. At the beginning stages the bots used basic rule-based evasion techniques which could not bypass advanced detection systems. The team implemented advanced adaptive algorithms that enabled cheat bots to alter their movement patterns and expansion rates based on the detection pressure they faced. The adjustment significantly enhanced the model's realism by better simulating real-world conditions where cheating strategies evolve to resist new detection techniques.

The team faced a major challenge in calibrating the cheat detection algorithm for optimal sensitivity. During the initial stages of development the algorithm showed an inability to correctly detect cheating bots while also generating too many false positive detections. The team developed a dynamic threshold-based scoring system that adjusted according to the changing behavior patterns of bots. Through adaptive thresholding the system achieved precise separation between genuine high-scoring players and cheating bots which reduced false positive detections.

Model Limitations & Areas for Improvement: The model produces functional and insightful simulation results but contains several inherent limitations. The model's primary limitation stems from its basic portrayal of player actions. Regular players in the simulation exhibit random movement and basic growth patterns while real-world players demonstrate strategic coordination during competitive play. The fidelity of the model to real-world gaming situations could improve through the implementation of complex human-like agent behavior.

The current model fails to capture complex social interactions between human participants. Legitimate players in actual gaming environments typically work together to spot cheaters which creates naturally occurring community-based moderation systems. The next version of the model should include player collaboration features that let multiple human agents work together to spot suspicious behaviors and report them to cheat detection systems to better represent community-focused anti-cheat methods.

The system responsible for scoring and flagging presents an area that requires enhancement. Adaptive cheating bots can sometimes avoid detection by keeping their scoring rates below set thresholds in existing threshold-based detection systems. The implementation of machine learning techniques that adaptively learn from changing cheat patterns may create a more resilient detection system while diminishing the capability of cheating agents to adapt.

Future Applications: These simulation results provide valuable insights for improving online platform governance strategies and advancing AI safety research. Our research provides valuable insights into creating adaptive cheat detection systems for gaming that respond to new cheating methods dynamically. Platform moderators alongside game developers could implement adaptive strategies demonstrated in our simulation to ensure fair play while avoiding excessive penalties to legitimate players.

The simulation framework is adaptable to multiple digital ecosystems that face ethical and operational problems related to AI-to-AI interactions beyond the gaming domain. Social media platforms that encounter automated content manipulation should implement adaptive detection systems comparable to those shown in this project. Detection frameworks in cybersecurity to identify AI-based cyberattacks can improve through adaptive and iterative methods demonstrated in this research.

Our research findings highlight the necessity for AI governance frameworks to be both transparent and capable of adaptation. The likelihood of false positives together with adaptive evasion directly impacts both the level of user trust and the credibility of the platform. Our simulation findings enable policymakers to develop balanced guidelines which ensure strong security alongside protection of user rights and fairness. Maintaining digital trust requires a crucial balance between security measures and countering evolving threats.

Going Above and Beyond: The reflective analysis identifies technical model challenges while proposing actionable improvements for future development. We support dynamic and adaptive detection mechanisms to help advance discussions about AI ethics and platform governance. The combination of community-driven detection systems with machine learning pattern recognition methods constitutes an innovative strategy which matches the ongoing transformation of AI interactions across media platforms.

References

Chen, M. (2024). AI cheating versus AI anti-cheating: A technological battle in game. Applied and Computational Engineering, 73(1), 222–227. https://doi.org/10.54254/2755-2721/73/20240402

Skinner, G., & Walmsley, T. (2019). Artificial intelligence and deep learning in video games: A brief review. In 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS) (pp. 404–408). IEEE. https://doi.org/10.1109/CCOMS.2019.8821783

Arai, K., Deguchi, H., & Matsui, H. (n.d.). Agent-based modeling meets gaming simulation. ResearchGate. https://www.researchgate.net/publication/321596055_Agent-Based_Modeling_Meets_Gaming_Simulation

Lehtonen, M. J., Vesa, M., & Harviainen, J. T. (2022). Games-as-a-disservice: Emergent value co-destruction in platform business models. Journal of Business Research, 141, 564–574. https://doi.org/10.1016/j.jbusres.2021.11.055

Jonnalagadda, A., Frosio, I., Schneider, S., McGuire, M., & Kim, J. (2021). Robust vision-based cheat detection in competitive gaming. Proceedings of the ACM on Computer Graphics and Interactive Techniques, 4(1), 1–18. https://doi.org/10.1145/3451259

Attestation

Our project team members unanimously agree that each person made substantial and unique contributions to both the final report and the project as a whole. The project team utilizes the Contributor Role Taxonomy (CRediT) to define each member's responsibilities and roles for transparent recognition of collaborative work.

Faraz Akbarzadeh (Leader/Facilitator):

- **Conceptualization:** The project initiator developed the concept and design framework to maintain vision consistency across project stages.
- **Project Administration:** The project timeline was managed while meetings were coordinated and team communication facilitated to maintain both progress and accountability.
- **Data Curation:** Managed simulation data organization and verification processes to achieve accurate and consistent analysis results for presentation purposes.

- **Writing – Original Draft (Ethical & Societal Reflections):** Created the first version of ethical considerations and societal implications by blending academic views and actual case studies.
- **Writing – Review & Editing:** I evaluated drafts critically while enhancing content clarity through refinement to improve coherence.

Adil Guluzade (Checker/Editor):

- **Visualization:** The creation of data visualizations through graphs and annotated visuals enabled better understanding of simulation results.
- **Writing – Original Draft (Observations & Results):** I created the analysis section where quantitative metrics were systematically presented alongside insights into emergent behavior patterns.
- **Review & Editing:** Verified that the report maintained linguistic accuracy while presenting coherent content with logical progression.
- **Validation:** Performed an extensive evaluation of simulation results to confirm their alignment with existing documentation and theoretical predictions.

Santusht Arora (Devil's Advocate/Software Developer):

- **Software Development:** The simulation model was developed through Mesa with optimization efforts to establish fundamental agent behaviors and dynamic interactions.
- **Methodology:** Provided significant input to the development of the methodological framework through the design of adaptive algorithms for cheat detection.
- **Validation:** Validation included testing simulation stability and accuracy while resolving computational difficulties and debugging problems.
- **Formal Analysis:** Our team performed an extensive analysis of emergent behaviors and recorded results together.

Collaborative Effort: The project demonstrated true teamwork because each team member engaged in developing both the concept and execution of the simulation and reporting process. The team exhibited unwavering dedication to quality work through their consistent engagement in discussions and problem-solving activities as well as iterative refinements. The attestation shows that the team maintained a unified commitment to thorough analysis and clear communication during all phases of the project.