Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-26)
Doctoral Consortium Track

Cybersecurity in AI-Enabled Digital Ecosystems: Evolutionary Game-Theoretic Modelling and Multi-Agent Defence

Adeela Bashir
School of Computing, Engineering and Digital Technologies, Teesside University
a.bashir@tees.ac.uk

Abstract

In this era of Artifical Intelligence (AI), cyber threats are no longer static, as modern attackers are intelligent, adaptive, and capable of exploiting vulnerabilities faster than traditional defences can respond. My research investigates how attack and defence strategies co-evolve in complex, adversarial environments by integrating evolutionary game theory (EGT) with AI-driven multi-agent systems (MAS). This combined framework allows me to model strategic interactions between competing actors, analyse how defensive decisions influence overall security outcomes, and identify the conditions under which defences hold or break down. A particular focus of my work is the emergence of coordinated behaviours in multi-agent AI systems that can compromise system reliability in ways that existing safety approaches do not adequately address. Through theoretical modelling and controlled multi-agent experiments, my research proposes mechanisms that reduce attack success and strengthen defensive resilience. This work contributes to the understanding of adaptive cyber conflicts and the development of more robust and trustworthy AI-enabled digital systems.

1 Introduction

Modern cyber threats are becoming increasingly adaptive and coordinated. Attackers now use automation, AI, and large-scale collaboration to exploit vulnerabilities in digital infrastructures [Mavikumbure *et al.*, 2024]. At the same time, defenders must protect complex environments such as industrial digital ecosystems, cloud services, and AI-enabled platforms while operating under limited resources and reducing cost [Mohamed, 2025].

Many traditional cybersecurity approaches rely on static risk models or classical game theory, which assume fixed strategies and perfectly rational agents. However, real cyber conflicts involve populations of attackers and defenders whose strategies evolve over time, influenced by incentives, observed successes, and environmental pressures [Skopik *et al.*, 2022; Sigmund, 2010]. EGT provides a natural framework for studying these dynamics because it captures strategy evolution at the population level and predicts future behaviour rather than one-shot rational decision-making [Smith and Price, 1973]. Moreover, recent advances in large language models (LLMs) and AI agents introduce multi-agent coordination and collusion risks, which traditional security frameworks are not designed to capture. Addressing these challenges requires a framework that moves from deterministic theoretical modelling to stochastic population analysis and empirical AI-system evaluation.

Motivation. Studying the evolution of attack and defence strategies in infinite and finite populations is necessary to appropriately model the real-world environments. Moreover, emerging AI systems introduce new security risks, including adversarial cooperation (collusion) among autonomous agents, which are not captured by existing models [Hammond *et al.*, 2025]. To solve these recent problems in AI and cyber security, this thesis develops a multi-layered framework for analysing and improving cybersecurity in AI-enabled systems by combining EGT and multi-agent AI experiments to understand adaptive attack–defence dynamics and design effective defence mechanisms.

The central research question of this thesis is:

RQ: How can evolutionary modelling and AI-based multi-agent systems be used to understand and mitigate adaptive cyber attacks in realistic and AI-driven environments?

2 Research Contributions

This thesis develops a progressive research programme consisting of three main contributions.

2.1 Co-evolutionary Dynamics in Infinite Populations

The first part models cybersecurity as a two-population evolutionary game between attackers and defenders [Bashir *et al.*, 2026a]. The model incorporates defence intensity, defence costs, attack benefits, and the probability of successful defence to make strategic decisions. Figure 1 illustrates the conceptual framework of our model within the NIST cybersecurity lifecycle. Using replicator dynamics and large-scale simulations of 100,000 randomly generated games, the analysis shows that systems with strong defence intensity tend to converge to stable low-attack environments, while weak defence environments remain vulnerable to persistent attacks.

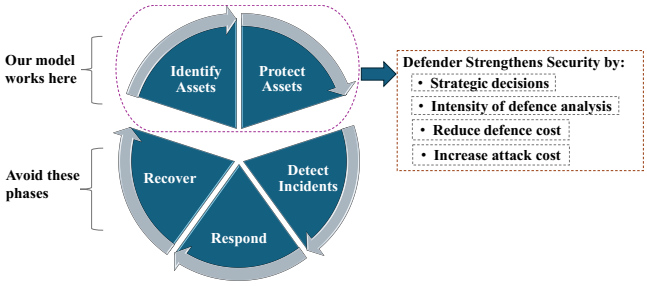

Figure 1: Conceptual framework of the proposed model within the NIST cybersecurity lifecycle, focusing on strategic defence decisions in the Identify and Protect phases.

Increasing the probability of defence success reduces the frequency of successful attacks and increases social welfare. The model is validated using a public dataset of cyber incidents $(2004 - 2020)$, demonstrating that the evolutionary dynamics align with real-world cyber conflict patterns.

2.2 Finite-Population Stochastic Analysis

The second contribution extends the model to finite populations [Bashir *et al.*, 2026b], which better represent realistic cyber ecosystems such as enterprise networks and IoT environments.

Using stochastic evolutionary dynamics with the Fermi update rule, fixation probabilities, and stationary distributions, the analysis shows that costly defence can discourage defenders from adopting strong protection, leading to attack-dominated outcomes. To address this challenge, the model introduces defence mechanisms using differential AI access and subsidy mechanisms and shows promising outcomes.

2.3 Multi-Agent LLM Collusion and Defence

The third contribution studies security risks in multi-agent AI systems based on LLMs [Bashir *et al.*, 2025]. This work studies the impact of communication among agents, as many AI decision systems rely on committees of agents that provide recommendations aggregated by a central decision agent.

Experiments in a clinical prescription task show that collusion among adversarial agents can turn the committee structure into an attack surface rather than a safeguard. Preliminary results show that the system accuracy collapses with the rise in the number of colluding agents.

To mitigate this risk, the research proposes a Verifier-Anchored Defence (VAD) architecture that validates decisions using external safety guidelines and restores accuracy. Experiments show that this mechanism eliminates collusion-induced harm with minimal computational overhead. This work highlights collusion as an underexplored threat in multi-agent AI systems and provides a practical defence strategy.

3 Real-world Implication

As AI systems become widely deployed and multi-agent LLM architectures grow in popularity, understanding security risks arising from strategic interactions among intelligent agents is both timely and important. This research provides insights that can guide the design of more resilient AI-enabled systems and inform safeguards for secure deployment of multi-agent AI technologies.

4 Future Research

The next stage of this research integrates reinforcement learning and federated learning with evolutionary game dynamics to model adaptive agents in adversarial multi-agent environments, while incorporating self-healing mechanisms to improve system resilience. This work aims to contribute to research in AI safety, adversarial learning, and cybersecurity.

Ethical Statement

Experiments are conducted in controlled environments using synthetic or public datasets, focusing on defensive applications and responsible experimentation without exposing real-world vulnerabilities.

References

[Bashir *et al.*, 2025] Adeela Bashir, Zia Ush Shamszaman, et al. Many-to-one adversarial consensus: Exposing multi-agent collusion risks in ai-based healthcare. *arXiv preprint arXiv:2512.03097*, 2025.

[Bashir *et al.*, 2026a] Adeela Bashir, Zia Ush Shamszaman, Zhao Song, and The Anh Han. Co-evolutionary dynamics of attack and defence in cybersecurity. *Knowledge-Based Systems*, 340, 2026.

[Bashir *et al.*, 2026b] Adeela Bashir, Zia Ush Shamszaman, Zhao Song, and The Anh Han. Strategic commitments shape collective cybersecurity under ai inequality. *arXiv preprint arXiv:2605.09415*, 2026.

[Hammond *et al.*, 2025] Lewis Hammond, Alan Chan, Jesse Clifton, Jason Hoelscher-Obermaier, Akbir Khan, Euan McLean, Chandler Smith, Wolfram Barfuss, Jakob Foerster, Tomáš Gavenčiak, et al. Multi-agent risks from advanced ai. *arXiv preprint arXiv:2502.14143*, 2025.

[Mavikumbure *et al.*, 2024] Harindra S Mavikumbure, Victor Cobilean, Chathurika S Wickramasinghe, Devin Drake, and Milos Manic. Generative ai in cyber security of cyber physical systems: Benefits and threats. In *2024 16th International Conference on Human System Interaction (HSI)*, pages 1–8. IEEE, 2024.

[Mohamed, 2025] Nachaat Mohamed. Artificial intelligence and machine learning in cybersecurity: a deep dive into state-of-the-art techniques and future paradigms. *Knowledge and Information Systems*, 67(8):6969–7055, 2025.

[Sigmund, 2010] Karl Sigmund. *The calculus of selfishness*. Princeton University Press, 2010.

[Skopik *et al.*, 2022] Florian Skopik, Markus Wurzenberger, and Max Landauer. Detecting unknown cyber security attacks through system behavior analysis. In *Cybersecurity of Digital Service Chains: Challenges, Methodologies, and Tools*, pages 103–119. Springer, 2022.

[Smith and Price, 1973] John Maynard Smith and George Robert Price. The logic of animal conflict. *Nature*, 246(5427):15–18, 1973.