Budapest University of Technology and Economics
Faculty of Electrical Engineering and Informatics
Department of Networked Systems and Services

Colonel Blotto in Cybersecurity: Real-World Data, Reinforcement Learning and Information Asymmetry

Scientific Students' Association Report

Author:

Zalán Nagy

Advisor:

Dr. Gergely Biczók

2025

Contents

Kivonat

A modern kiberbiztonsági környezetben a szoftversérülékenységek gyors felfedezése és kihasználása állandó versenyhelyzetet teremt a támadók és a védők között. A gyártói ökoszisztémák kitettsége napról napra változik, mivel az újonnan közzétett sérülékenységek súlyossága és kihasználhatósága folyamatosan alakítja a globális kockázati felületet. Ebben a környezetben a korlátozott védekezési és támadási erőforrások elosztása központi stratégiai kérdéssé válik.

A dolgozat ezeket az adaptív kölcsönhatásokat egy valós sérülékenységi adatokra épülő, dinamikus Colonel Blotto keretrendszerrel modellezi. A csatamezők szoftvergyártókat jelentenek, amelyek napi értéke a CVSS súlyosság és az EPSS kihasználhatóság szorzataként kerül meghatározásra. Ez a mutató egyaránt tükrözi a várható hatást és a támadási valószínűséget. Két megerősítéses tanuláson alapuló ügynök, egy támadó és egy védő, ismételten elosztja korlátozott erőforrásait a csatamezők között, majd a tapasztalatok alapján módosítja stratégiáját. A modell két időléptékben működik, a rövid távú tanulási epizódokat napi adatfrissítések egészítik ki, ezáltal összekapcsolva a taktikai alkalmazkodást a hosszabb távú környezeti változásokkal.

A vizsgálat négy információs aszimmetria rezsimet különböztet meg, amelyek eltérő láthatósági és hírszerzési feltételeket írnak le. A januári, májusi és augusztusi adathalmazokon végzett kísérletek azt mutatták, hogy a rendszer erősen időben stabil, és a védő minden esetben felülmúlta a támadót. A védői hírszerzéssel rendelkező szcenáriók bizonyultak a leghatékonyabbnak, amelyekben a támadások sikeressége jelentősen csökkent, miközben a legkritikusabb gyártók védelmi lefedettsége közel teljes volt. A támadói információelőnyt tartalmazó helyzetek ugyan növelték az átmeneti támadási nyereséget, de nagyobb ingadozással és gyengébb hosszú távú stabilitással jártak. Az entrópia alapú mérések azt mutatták, hogy mindkét ügynök adaptív és nem konvergens tanulási mintázatot követett, amely statikus egyensúly helyett dinamikus egyensúlyt eredményezett.

Az eredmények alapján a megerősítéses tanuláson alapuló adatvezérelt védelem tartósan felülmúlja a statikus erőforrás elosztási megközelítéseket, és hatékonyan alkalmazkodik a változó kockázati környezethez. A naprakész sérülékenységi információkhoz való hozzáférés gyorsítja a tanulási folyamatot és növeli a védelmi hatékonyságot, ami gyakorlati iránymutatást ad a patch menedzsment és az NIS2 előírásainak megfelelő kockázatkezelési gyakorlatokhoz. A dolgozat egy reprodukálható, adatvezérelt megközelítést mutat be, amely alkalmas a támadó és védő ügynökök adaptív viselkedésének vizsgálatára a folyamatosan változó kiberfenyegetési környezetben.

Kulcsszavak: kiberbiztonság, játékelmélet, Colonel Blotto, erőforrás-elosztás, megerősítéses tanulás, CVE, CVSS, EPSS.

Abstract

In the modern cybersecurity landscape, the rapid discovery and exploitation of software vulnerabilities create a continuous competition between attackers and defenders. The exposure of vendor ecosystems changes every day as newly disclosed vulnerabilities reshape the global risk surface through their severity and exploitability. Within this environment, the allocation of limited defensive and offensive resources becomes a central strategic challenge.

This thesis models these adaptive interactions through a dynamic Colonel Blotto framework based on real vulnerability data. The battlefields represent software vendors whose daily value is determined by the product of CVSS severity and EPSS exploitability, reflecting both impact and likelihood of attack. Two agents, an attacker and a defender, repeatedly distribute their limited resources across the battlefields and update their strategies using reinforcement learning with Q learning. The model operates on two time scales where short term learning episodes are complemented by daily data updates that connect tactical adaptation with longer term environmental change.

The study distinguishes four information asymmetry regimes, each representing different levels of visibility and intelligence access between the agents. Experiments conducted across datasets from January, May, and August demonstrated strong temporal robustness, with the defender consistently outperforming the attacker. Scenarios with defender-side intelligence produced the most effective protection, substantially reducing breach probability while maintaining near-complete coverage of the most critical vendors. Situations with attacker-side information advantage increased short-term gains but resulted in greater fluctuation and weaker long-term stability. Entropy-based analyses showed that both agents followed adaptive and non-convergent learning patterns, leading to a dynamic rather than a static equilibrium.

The results show that reinforcement learning based, data driven defense consistently outperforms static allocation methods and adapts effectively to changing risk environments. Access to timely vulnerability intelligence accelerates the learning process and increases protection efficiency, offering practical guidance for patch management and for risk management processes aligned with the NIS2 directive. The thesis presents a reproducible, data-driven approach suitable for analyzing the adaptive behavior of attacker and defender agents in an ever-changing cybersecurity environment.

Keywords: cybersecurity, game theory, Colonel Blotto, resource allocation, reinforcement learning, CVE, CVSS, EPSS.

Chapter 1

Introduction

1.1 Motivation

1.1.1 The strategic nature of cyberspace

During the past decades, cyberspace has transformed from a technical infrastructure into a strategic environment comparable to the traditional domains of land, sea, air, and space. Modern societies depend on digital systems to operate critical services such as finance, transportation, healthcare, energy, and communication. Because these systems are interconnected, the compromise of a single component can spread, leading to systemic disruptions. Consequently, cyberspace has become an arena of continuous contestation between offensive and defensive actors whose actions are motivated not only by technical curiosity or criminal intent but also by economic and geopolitical incentives [1, 2].

Cyber operations today extend beyond data theft or service disruption. They encompass espionage, disinformation, industrial sabotage, and strategic coercion. In this environment, every vulnerability represents a potential foothold for exploitation, and every patch or security control corresponds to an investment in resilience. The frequency, sophistication, and scale of cyber incidents illustrate that the cyber domain functions as an ongoing, multi-agent game where decisions have measurable payoffs and costs.

1.1.2 The economics of cyberattacks

From an economic perspective, a cyberattack can be interpreted as an investment decision. An attacker dedicates time, computing power, and skilled personnel to finding, weaponizing, and using exploits, aiming for returns such as access to data, financial profit, or strategic leverage. Just like in financial markets, the attacker faces an uncertain return on investment, as the exploit may be patched, detected, or yield lower value than expected. Rational attackers therefore seek to maximize expected profit while minimizing operational cost and detection risk.

Defenders, in contrast, face a cost-minimization problem. They cannot protect every system equally because budgets, expertise, and attention are limited. Each defensive action, such as patching a vulnerability, investing in monitoring, or conducting incident response, involves an opportunity cost that diverts resources from other potential defenses. Effective protection therefore depends on clear prioritization, which means determining which vulnerabilities, assets, or vendors deserve the most attention [3, 4]. This prioritization represents a resource-allocation problem under uncertainty.

1.1.3 Vulnerabilities as drivers of risk

At the core of every strategic interaction in cyberspace lie software vulnerabilities. These are design or implementation flaws that allow breaches of confidentiality, integrity or availability. The life cycle of a vulnerability, which extends from its discovery and public disclosure to patching and possible exploitation, creates a time frame in which attackers and defenders compete for control.

The global management of vulnerabilities is supported by standardized databases such as the Common Vulnerabilities and Exposures (CVE), which provides unique identifiers and textual descriptions for known issues. However, not all vulnerabilities pose equal risk. To quantify severity, the Common Vulnerability Scoring System (CVSS) expresses each CVE on a numerical scale from 0 to 10, combining exploit complexity, impact, and scope. Still, CVSS alone describes only potential impact, not the likelihood of real-world exploitation. For this reason, the Exploit Prediction Scoring System (EPSS) has been developed to estimate the probability that a vulnerability will be exploited in the near future based on historical data [5, 6, 7].

Together, these metrics define a continuously evolving risk landscape. As new CVEs are published daily and EPSS scores fluctuate based on threat-intelligence updates, the perceived value of defending or attacking a given software ecosystem changes dynamically. This constant evolution motivates the need for adaptive defense strategies capable of responding to new information in near real time.

1.1.4 Attackers and defenders as adaptive decision makers

Both attackers and defenders operate under uncertainty and incomplete information. Attackers rarely know whether a target has been patched or whether defensive monitoring is active, while defenders cannot predict which vulnerabilities attackers will focus on. In many cases, attackers possess a structural or temporal advantage—often referred to as the attacker's advantage, which arises from their ability to act first, exploit newly discovered vulnerabilities, and select the timing and method of attack before defenders can respond [8, 9, 10, 11]. Despite this uncertainty, both sides behave strategically and gradually adapt over time. Attackers modify their tactics when detection rates rise, and defenders revise their patching priorities when new exploits appear.

These adaptation processes can be interpreted as learning phenomena. Attackers learn from successful intrusions which techniques are effective, and defenders learn from incidents and threat intelligence which assets need stronger protection. Each side observes signals such as exploit attempts, vulnerability disclosures or intrusion detections and updates its beliefs based on these observations. This feedback cycle closely follows the principles of reinforcement learning (RL), where agents improve their policies through repeated interaction with the environment and the rewards or penalties that result from their actions [12, 13].

1.1.5 Rationality and decision criteria

Although not all actions in cyberspace are entirely rational, assuming bounded rationality provides a clear foundation for analysis. In this context, bounded rationality does not refer to human cognitive limitations but to the computational and informational constraints of algorithmic agents. Each actor optimizes its objective function based on partial observations, finite learning capacity, and limited feedback.

- The attacker aims to maximize the expected value of successful intrusions while taking into account the costs of operation and the probabilities of success and detection.
- The defender aims to minimize the expected loss by balancing the cost of protection against the potential damage caused by compromise.

This alignment of incentives between profit and cost naturally leads to a game theoretic interpretation. Each player's optimal decision depends on the choices of the other, creating a strategic equilibrium that can evolve as information and resources change. Classical equilibrium concepts assume fixed payoffs and perfect rationality, but these conditions rarely exist in cybersecurity. Real environments change over time, and learning occurs gradually rather than instantaneously. Dynamic models based on learning are therefore necessary to approximate realistic behavior.

1.1.6 From cyber conflict to the Colonel Blotto analogy

The Colonel Blotto game provides an intuitive abstraction of the problem of resource allocation. In its original version two military commanders divide a limited number of troops across several battlefields, and each battlefield is won by the player who assigns more units. The key idea is that victory depends not only on the total resources but also on the way they are distributed. In the context of cybersecurity the analogy can be expressed as follows:

- Battlefields represent software vendors or technology domains.
- Troops represent limited offensive or defensive resources.
- Victory on a battlefield corresponds to a successful exploitation or defense.

The analogy can be extended in three important ways. First, the cyber battlefields are not static because their values change over time as new vulnerabilities appear. Second, the information available to the players is asymmetric because attackers may have intelligence leaks or exploit repositories, while defenders may rely on vulnerability databases or early warnings. Third, learning takes place continuously as both sides refine their strategies through experience [8, 9]. To capture these properties the Colonel Blotto framework is combined with reinforcement learning and real world vulnerability data [14, 15, 16, 17].

1.2 Problem Statement

The research presented in this study focuses on understanding how learning and information asymmetry shape the evolution of strategies between attackers and defenders in a dynamic cyber environment.

Cyber defense can be viewed as a sequential decision making process under uncertainty, in which both sides adapt based on partial observations of the other's behavior. To model this interaction the thesis proposes a dynamic data driven Colonel Blotto game where each battlefield corresponds to a software vendor whose exposure level changes daily as a function of real vulnerability data.

1.2.1 Dynamic structure of the model

The model captures cyber conflict dynamics on two interconnected time scales. On the inner level, attacker and defender agents repeatedly interact through reinforcement-learning episodes, adjusting their allocation strategies based on observed outcomes. On the outer level, the environment itself evolves as vendor-level risk values change over time according to real CVSS and EPSS data.

This dual structure links short-term adaptive learning with long-term shifts in the vulnerability landscape, allowing the simulation to reflect how agents continuously readjust to new information.

1.2.2 Information asymmetry and rational interaction

In practical cybersecurity perfect information is unattainable. Attackers and defenders rarely share the same visibility over vulnerabilities or each other's actions. This asymmetry fundamentally influences decision making because an attacker with insider knowledge can prioritize high value targets, while a defender with up to date vulnerability intelligence can allocate resources more efficiently [18, 19].

To capture these effects systematically the model defines several information regimes $R0$ to $R3$ that differ in what each player can observe.

- access to true or estimated vulnerability values based on EPSS visibility,
- observation of the opponent's recent strategies through strategic visibility.

By analyzing these configurations the study examines how unequal access to information changes both the learning trajectories and the eventual equilibrium behavior of the players. The model also explores whether more accurate or timely information leads to faster adaptation and higher effectiveness, or whether adaptation alone can compensate for informational disadvantage.

1.2.3 Rationale for using reinforcement learning

Traditional equilibrium-based game theory assumes perfectly rational players operating under static payoffs. Cybersecurity interactions, however, are iterative, uncertain, and shaped by incomplete information. Reinforcement learning (RL) provides a more realistic framework for modeling such adaptive behavior [20, 21] as agents refine their strategies through experience rather than analytical optimization. This approach captures the trial-and-error adaptation typical of real defensive operations and adversarial campaigns. The technical formulation of the RL agents and learning algorithms is described in Section 2.3.

1.2.4 Practical interpretation

Although abstract, this model represents a real operational challenge faced by organizations when deciding how to distribute finite cybersecurity resources across competing priorities. Patching every vulnerability immediately is infeasible, and scanning every vendor continuously is unrealistic. Similarly, attackers cannot exploit every target simultaneously. Both sides must prioritize, reallocate and learn from the outcomes. By grounding

this allocation process in measurable data such as CVSS and EPSS and in learning algorithms such as Q learning, the model creates a controlled yet realistic environment to test hypotheses about adaptive behavior under information asymmetry.

It also connects theoretical constructs with practical cybersecurity management by emphasizing the importance of continuous feedback, situational awareness and adaptive learning in achieving effective defense.

The central research question addressed in this thesis can therefore be summarized as follows.

> **How does information asymmetry affect the learning dynamics and overall performance of attacker and defender agents in a data driven Colonel Blotto game?**

1.2.5 Scope and focus

The scope of the model deliberately abstracts away from low level technical details such as specific exploit mechanisms. Instead it focuses on the strategic layer of cybersecurity that involves allocating attention and effort across multiple sources of risk. This abstraction allows for generalizable insights into how learning and access to intelligence shape defensive efficiency. The model's results can be applied to practical areas including patch management prioritization, calibration of vulnerability scores and planning of resources for security operations centers.

1.3 Research Objectives

To answer the research question the study pursues the following objectives.

RO1: Formulate a dynamic Colonel Blotto model where battlefields evolve from real time CVE, CVSS and EPSS data.

RO2: Implement attacker and defender agents using Q learning with configurable visibility parameters that represent different intelligence conditions.

RO3: Define four information asymmetry scenarios $R0$ to $R3$ that capture various combinations of data availability and strategic observability.

RO4: Evaluate the agents' learning dynamics and outcomes under these scenarios in order to understand how information asymmetry affects adaptation and strategic variability.

1.4 Contributions

This thesis provides a data-driven, learning-based perspective on strategic interactions in cybersecurity. It models attacker–defender adaptation as a reinforcement-learning process within a dynamic Colonel Blotto framework grounded in real vulnerability data.

The first contribution lies in the conceptual formulation of the model, presented in Chapter 3. Here, a dynamic Colonel Blotto structure is introduced in which each battlefield

represents a real software vendor, and its value evolves daily based on CVSS and EPSS data. This formalization connects theoretical game-theoretic constructs with measurable indicators of cyber risk, providing a bridge between abstract strategy design and operational vulnerability management.

The second contribution concerns the technical implementation of the simulation framework, detailed in Chapter 4. A modular reinforcement-learning environment was developed, integrating Q-learning agents and multiple information-asymmetry regimes. This architecture enables controlled experimentation on how bounded-rational agents adapt their allocation strategies under varying visibility and intelligence conditions.

The third contribution consists of the empirical findings and their interpretation, discussed in Chapter 5. Through extensive simulations, the study demonstrates how information asymmetry and adaptive learning jointly shape strategic behavior. The results show that access to vulnerability intelligence substantially improves defensive efficiency, while the absence of perfect information leads to oscillatory yet stable adaptive dynamics that reflect real-world cyber competition.

Finally, Chapter 6 summarizes the broader implications of these findings and outlines potential directions for future research, including deep-reinforcement extensions, graph-based battlefield modeling, and live data integration. Overall, the thesis bridges theoretical game-theoretic models and empirical cyber-risk analytics, demonstrating that adaptive, data-driven defense can measurably enhance protection in evolving threat environments.

Chapter 2

Background and Related Work

2.1 Colonel Blotto Games in Security

The Colonel Blotto game, first introduced by Émile Borel in 1921 and later formalized by Roberson [14], is a classical resource-allocation game that models competitive decision-making under scarcity. Two rational players, typically referred to as Colonel A and Colonel B, each possess a finite budget of resources that must be distributed simultaneously across a fixed number of battlefields [14, 16]. Each battlefield represents an independent contest, where the player allocating the greater share wins that battlefield, and the overall payoff is determined by the total number (or weighted sum) of battlefields won. The game's zero-sum nature ensures that one player's gain is the other's loss, making it an archetype of strategic conflict.

Mathematically, the Colonel Blotto game belongs to the broader class of mixed-strategy equilibria in continuous action spaces. Analytical solutions exist only for highly symmetric cases, in which all battlefields carry equal value and the players possess equal budgets. In these classical settings, equilibrium strategies are probabilistic resource distributions that balance expected payoffs across fronts, preventing the opponent from exploiting predictable allocations. For asymmetric or weighted variants, where the battlefields have different importance and the players possess unequal resources, closed-form equilibria generally do not exist. This limitation motivates the use of computational and simulation-based approaches.

Beyond its original military interpretation, the Colonel Blotto model has been applied in several fields. It appears in political campaigns, in marketing and research competition, and in network security games. Its main idea is the division of limited resources among several parallel contests. This principle captures the essence of strategic trade-offs in any system where attention, funding, or computing power must be shared between competing goals.

In cybersecurity, this abstraction translates naturally into the attacker–defender dilemma. A defender must decide how to allocate limited defensive capacity (e.g., monitoring, patching, or hardening) across multiple assets, while an attacker decides how to distribute offensive effort toward potential targets. This simultaneous decision process mirrors the structure of the Colonel Blotto game. Each digital asset can be seen as a battlefield, and each side's allocation determines the outcome of that particular engagement. If both sides could observe each other's allocations, the game would be trivial, instead, uncertainty and incomplete information make strategic randomization essential.

Early applications of the Blotto framework to cyber defense include the work of Chia and Chuang [15], who modeled phishing campaigns as multi-front contests between attackers' resources and defenders' user-training attention. Later studies extended the model by introducing heterogeneous target values, variable attacker costs, and limited information. These developments highlight that cyber conflicts are inherently non-uniform. Attackers tend to focus their resources on weak but valuable targets, while defenders must distribute their limited budgets across assets with different levels of exposure and importance.

Modern extensions of the Blotto game increasingly address stochasticity, incomplete information, and temporal change [16, 17, 22]. Dynamic variants allow players to reallocate resources over time, representing evolving security landscapes in which new vulnerabilities emerge and old ones are patched. Although these models remain analytically challenging, they provide a flexible foundation for studying adaptive defense and attack behaviors under resource constraints, an essential step toward the data-driven, learning-based approaches developed later in this thesis.

2.2 Cybersecurity Risk Metrics

Quantitative risk metrics form the empirical backbone of modern vulnerability management. They provide standardized, machine-readable indicators that allow security teams and automated systems to prioritize defensive actions consistently across software ecosystems. Among these, the Common Vulnerabilities and Exposures (CVE), the Common Vulnerability Scoring System (CVSS), and the Exploit Prediction Scoring System (EPSS) represent the three most important pillars. Together, they define how vulnerabilities are enumerated, how their technical severity is measured, and how their real-world exploitation likelihood is estimated.

2.2.1 CVE

The Common Vulnerabilities and Exposures (CVE) system, launched in 1999 and maintained by the MITRE Corporation, provides a unified naming convention for publicly disclosed software vulnerabilities. Each CVE entry corresponds to a unique identifier (e.g., CVE-2025-12345) that links advisories, patches, and exploit data across vendors and databases. The goal of CVE is not to describe vulnerabilities in depth, but to serve as an indexing layer and effectively creating a shared lexicon for vulnerability intelligence [5].

Every CVE entry contains a brief summary, affected products (via the Common Platform Enumeration, CPE), and external references. The CVE data is consumed by the U.S. National Vulnerability Database (NVD), which enriches it with structured metadata including CVSS vectors, vendor tags, and publication timestamps. As of 2025, over 240,000 unique CVEs have been published, with roughly 30,000 new entries each year. This explosion underscores both the efficiency of modern vulnerability disclosure pipelines and the difficulty defenders face when prioritizing patching efforts.

While the CVE system has greatly improved coordination, it remains descriptive rather than evaluative. It assigns identifiers but does not indicate how dangerous a vulnerability is or how likely it is to be exploited. For this reason, the CVSS and EPSS frameworks are necessary.

Figure 2.1: Example CVE entry (CVE-2025-53792) illustrating the structure of a typical vulnerability record in the National Vulnerability Database (NVD).

2.2.2 CVSS

The Common Vulnerability Scoring System (CVSS) provides a standardized framework for assessing the intrinsic technical severity of vulnerabilities [6]. Developed under the Forum of Incident Response and Security Teams (FIRST) and standardized by ITU-T as Recommendation X.1521, CVSS decomposes vulnerability impact into three metric groups:

- **Base metrics:** intrinsic properties of a vulnerability that remain constant over time, such as attack vector, attack complexity, privileges required, user interaction, and impacts on confidentiality, integrity, and availability.
- **Temporal metrics:** properties that may change over the vulnerability lifecycle, such as exploit code maturity, remediation level, and report confidence.
- **Environmental metrics:** contextual adjustments reflecting the importance of the affected system and local security requirements.

The most widely used component is the Base Score, computed on a logarithmic 0 to 10 scale. CVSSv2 and CVSSv3 differ mainly in granularity and normalization. For example, the CVSSv3.1 Base formula defines the Exploitability subscore as a multiplicative function of attack vector, complexity, privileges, and user interaction, while the Impact subscore measures the potential damage to the CIA triad, which describes the three core security principles of confidentiality, integrity, and availability.

Despite its rigor, the CVSS score has been frequently misinterpreted as a proxy for risk, rather than for intrinsic severity. As Jacobs et al. note in the EPSS study [7], this misinterpretation can lead to inefficient remediation. Organizations often patch high-CVSS vulnerabilities that are unlikely to be exploited, while leaving genuinely dangerous low-CVSS vulnerabilities unattended. Empirical analyses have confirmed that CVSS-based prioritization is often no better than random selection when measured against actual exploit data in the wild [7, 23, 24]. This discrepancy arises because CVSS does not consider attacker incentives, exploit availability, or market dynamics, it is a static technical measure detached from real threat activity.

Nonetheless, CVSS remains indispensable. Its structured vector representation allows interoperability across tools and serves as the mathematical foundation for hybrid metrics

like EPSS. In practice, many organizations use CVSS thresholds (e.g., CVSS ≥ 7.0) to trigger patching Service-Level Agreements (SLAs) or compliance mandates such as PCI-DSS and NIST directives. However, the growing consensus in both academia and industry is that CVSS must be complemented, not replaced, by probabilistic models that incorporate real exploitation evidence.

2.2.3 EPSS

The Exploit Prediction Scoring System (EPSS) was proposed by Jacobs, Romanosky, Edwards, Adjerid, and Roytman [7] as the first open, data-driven framework for estimating the probability that a vulnerability will be exploited in the wild within 12 months of public disclosure. While CVSS quantifies the potential impact of exploitation, EPSS focuses on its likelihood, thereby complementing the deterministic perspective of CVSS with an empirical, probabilistic one.

EPSS is grounded in empirical machine learning. A logistic regression model trained on over 25,000 vulnerabilities disclosed between 2016 and 2018. The authors aggregated data from multiple sources, including MITRE's CVE repository [25], NIST's NVD [26], and exploit databases such as Exploit-DB [27], Metasploit [28], D2 Elliot [29], and Canvas [30], as well as telemetry feeds from security vendors such as Proofpoint [31], Fortinet [32], and AlienVault [33].

Of these vulnerabilities, only about 3.7% were observed to be exploited in the wild. The model identified sixteen predictive features, including vendor identity (e.g., Microsoft, Adobe, IBM, HP), exploit code availability (proof-of-concept or weaponized), and descriptive keywords such as "remote," "code execution," or "denial of service."

A key design goal of EPSS is transparency. Instead of complex neural architectures, it employs a regularized logistic regression model with a concise and interpretable feature set. This design enables practitioners to understand why a particular vulnerability receives a high score (for instance, if exploit code is publicly available for a widely used product) and to compute the score using openly published parameters. EPSS values are normalized to a probability in the range [0,1], directly interpretable as the likelihood of exploitation within one year. Because the model coefficients and input features are public, EPSS can be continuously retrained as new data are released.

Conceptually, EPSS bridges the gap between vulnerability severity and threat. It introduces a probabilistic layer that estimates the likelihood of real-world exploitation, providing a dynamic complement to static severity metrics. The EPSS dataset is updated daily, integrating new CVE disclosures, exploit code availability, and telemetry-based observations from real-world attack activity. By 2025, it encompasses over 200,000 scored vulnerabilities, providing a rich, real-time empirical foundation for research in cyber-risk assessment, prioritization, and adaptive defense modeling.

2.2.4 Combined Use and Limitations

Combining CVSS and EPSS provides a unified measure that integrates severity and exploitability.

While this composite perspective strengthens quantitative cyber-risk assessment, both components have inherent limitations. CVSS relies on consistent expert evaluation of vector parameters, which may vary across organizations. EPSS, though data-driven, depends on the visibility of exploitation telemetry and may underrepresent targeted or stealthy

attacks that remain outside public datasets. Nevertheless, the two together offer the most transparent and publicly available basis for modeling adaptive attacker–defender dynamics in reinforcement learning and game-theoretic contexts.

2.3 Reinforcement Learning

Reinforcement Learning (RL) provides a computational mechanism for agents to discover optimal policies through repeated interaction with an environment [12]. An RL agent observes a state, selects an action, receives a reward, and updates its policy to maximize expected cumulative return. Among the simplest and most widely used algorithms is tabular Q-learning, which iteratively refines a value estimate for each state–action pair. RL agents face the well-known exploration–exploitation trade-off, often handled through an ε-greedy policy or its decaying variants, which balance random exploration and reward-driven exploitation [13, 34, 35].

In the context of Colonel Blotto games, RL offers a means of approximating equilibria in environments where analytic solutions are difficult to obtain or the payoff landscape changes over time. Noel [20] implemented Q-learning agents competing in multi-battlefield Blotto scenarios and showed that learned strategies consistently outperform random baselines. Building on this foundation, An and Zhou [36] proposed a deep reinforcement learning framework for game-theoretic resource allocation on graphs, extending Blotto games to spatially constrained domains using the Deep Q-Network (DQN) and Proximal Policy Optimization (PPO) algorithms. Their agents adapted resource movements across connected nodes, achieving balanced outcomes even on asymmetric topologies. These works demonstrate that RL agents can autonomously learn nuanced allocation heuristics, mirroring the adaptive decision-making observed in real cyber operations.

2.4 Related Work

Game-theoretic modeling of cybersecurity has evolved from abstract deterrence theory to empirically grounded, learning-based simulations. Early conceptual works by Arquilla and Ronfeldt [1], Rid [2], and Moore et al. [18] explored the strategic and ethical dimensions of cyber conflict, emphasizing trade-offs between offensive and defensive postures. Subsequent quantitative models examined economic incentives, information asymmetry, and the role of uncertainty in cyber defense [4, 3]. Ransomware and attack-insurance models [37, 38] further demonstrated how payoff structures influence optimal defensive spending.

Recent research increasingly couples such game-theoretic formalisms with learning components. Noel's RL-based Blotto framework [20] introduced adaptive agents capable of evolving their strategies through self-play, while An and Zhou's GRAG model [36] generalized this to multi-step graph-constrained games. These developments converge toward a paradigm in which cybersecurity strategy emerges from continual learning rather than static optimization.

Data-driven parameterization has also gained traction. Mucsi's BME thesis [39] employed CVE- and CVSS-based battlefield valuations to ground Blotto simulations in observable risk data. Complementary empirical analyses confirm the predictive relevance of such metrics for real-world breaches. Integrating these data sources with adaptive RL agents enables realistic modeling of how defenders and attackers respond to evolving vulnerability landscapes.

In summary, prior research establishes three converging lines of inquiry: (i) strategic allocation frameworks derived from the Colonel Blotto game, (ii) quantitative risk metrics linking vulnerabilities to exploitation likelihood, and (iii) learning-based adaptation mechanisms for sequential decision-making. The present thesis unifies these strands by developing a reinforcement-learning-driven Colonel Blotto model parameterized with real-world CVSS and EPSS data, thereby bridging theoretical game models and practical cybersecurity analytics.

Additionally, regulatory frameworks such as the EU *NIS2 Directive* define mandatory cyber risk-management and incident-reporting requirements for operators of essential and important entities, reinforcing the practical relevance of adaptive, data-driven defense allocation [40].

Chapter 3

Model Design

3.1 Conceptual Overview

In the proposed framework, cybersecurity is represented as a continuous strategic interaction between an attacker and a defender competing over a set of vulnerable software ecosystems. Each ecosystem (or battlefield) corresponds to a software vendor whose systems expose measurable risk due to known vulnerabilities. This abstraction enables the study of how rational decision-making and limited information jointly shape the dynamics of cyber conflicts.

At the core of this formulation lies the classical definition of cyber risk as the product of likelihood and impact [6, 5, 7].

$$\text{Risk}_k = \text{Likelihood}_k \times \text{Impact}_k = \text{EPSS}_k \times \text{CVSS}_k, \tag{3.1}$$

where EPSS_k captures the probability of exploitation for vendor k, and CVSS_k expresses the potential impact if an exploit occurs. The resulting combined score Risk_k therefore quantifies the expected loss or gain associated with targeting or protecting that vendor. A high Risk_k corresponds to a battlefield that is both likely to be attacked and costly to compromise, which makes it a prime focus for both offensive and defensive resource allocation.

These risk values define the relative attractiveness of each battlefield for both players. A higher Risk_k represents a more valuable target for the attacker and a more critical asset for the defender. Consequently, the players' strategies can be interpreted as risk-weighted resource allocations that evolve dynamically as the underlying vulnerability data changes over time.

The Colonel Blotto formulation is particularly suitable for cybersecurity because it captures the essence of strategic trade-offs under resource constraints [14, 15, 16]. Unlike single-target defense models, Blotto explicitly represents the interdependence between multiple assets and the necessity to distribute finite resources among them.

This conceptual structure is illustrated in Figure 3.1, which depicts the attacker and defender distributing their resources across multiple battlefields that represent software vendors with varying risk levels.

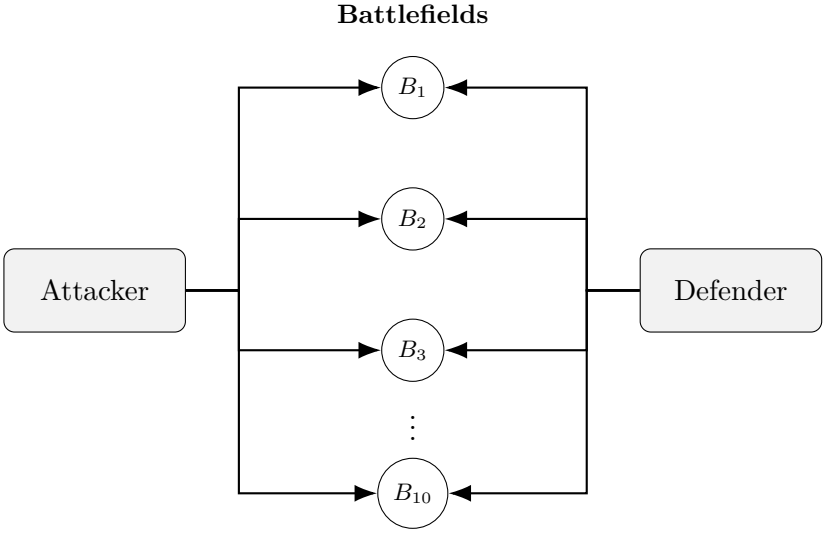

Figure 3.1: Conceptual overview of the attacker–defender interaction across multiple battlefields.

3.2 Game Setup

The model extends the classical Colonel Blotto framework into a dynamic, data-driven cybersecurity environment. Ten battlefields are defined, each representing a software vendor with the highest number of newly disclosed vulnerabilities during the observation period. The value of each battlefield is derived from its aggregated risk exposure:

$$v_k = \mathrm{CVSS}_k \cdot \mathrm{EPSS}_k. \tag{3.2}$$

Each v_k value is normalized using the min–max method, which scales the data between the minimum and maximum values while preserving the relative differences between them. This normalization ensures comparability across vendors and time periods, preventing any single vendor from dominating the results purely due to scale differences.

The dynamic update mechanism ensures that the model remains grounded in real-world data. By recalculating battlefield values on a daily basis, the simulation reflects the temporal patterns of vulnerability discovery, disclosure, and patching observed in public datasets such as the National Vulnerability Database (NVD) and the Exploit Prediction Scoring System (EPSS) feed [5, 7]. This temporal granularity enables the analysis of how learning stability and adaptation efficiency are influenced by the natural dynamics of the cyber-risk landscape.

3.3 Players and Learning Mechanism

Two reinforcement-learning (RL) agents participate in the game: an attacker and a defender. Both repeatedly allocate limited resources across the battlefields and update their strategies based on the observed payoffs after each round.

- **Attacker:** Seeks to maximize cumulative exploited value. Resources correspond to time, exploits, or operational effort invested into attack campaigns. The attacker learns which vendors are more profitable to target based on historical outcomes.
- **Defender:** Seeks to minimize total residual risk by allocating inspection and patching capacity. The defender learns which assets require stronger protection to minimize cumulative loss over time.

Both agents follow an ε-greedy exploration strategy within the Q-learning algorithm, maintaining a balance between exploring new strategies and exploiting the best-performing ones discovered so far. Learning rates (α) and exploration parameters (ε) are held constant to ensure fair comparison across different simulation scenarios.

In contrast to traditional game-theoretic equilibria, where players are assumed to possess complete knowledge and infinite reasoning capacity, reinforcement learning models bounded rationality more realistically. Agents do not optimize analytically but gradually approximate effective strategies through repeated experience. This learning-based approach mirrors real-world cyber operations, where defenders improve their posture through feedback and attackers evolve tactics based on observed defenses and discovered vulnerabilities [8, 21].

3.4 Payoff Function

The reward in each episode is computed by comparing the attacker's and defender's allocations across battlefields. If the attacker invests more resources on battlefield k, that field is considered compromised; otherwise, the defender maintains control. The defender's reward equals the sum of the values of successfully defended battlefields, while the attacker's reward is complementary:

$$R^{(\text{att})} = \sum_{k=1}^{N} \mathbb{I}(a_k > d_k)\, v_k, \quad R^{(\text{def})} = \sum_{k=1}^{N} \mathbb{I}(d_k \geq a_k)\, v_k, \tag{3.3}$$

where a_k and d_k denote the respective resource allocations of the attacker and defender on battlefield k, and $\mathbb{I}(\cdot)$ is the indicator function.

To increase realism, the zero-sum assumption is relaxed using a probabilistic success factor, which introduces stochasticity into the outcome. This term represents uncertainties such as exploit reliability, network heterogeneity, human error, or random environmental noise, factors that commonly influence the success of both cyberattacks and defensive interventions. This stochastic component makes the simulation less idealized and allows convergence analyses that reflect noisy, imperfect decision environments [3, 4, 17].

3.5 Information Scenarios

Information asymmetry is a central dimension of the model. Four configurations (R_0–R_3) are defined to simulate different combinations of data visibility and strategic observability [19].

Table 3.1: Information-asymmetry configurations defining what each agent can observe.

Scenario	EPSS visibility		Strategy observability	
	Attacker	Defender	Attacker	Defender
R_0	✓	✓	✗	✗
R_1	✗	✓	✗	✓
R_2	✓	✗	✗	✗
R_3	✓	✓	✗	✓

Each regime captures a realistic operational setting, for example, R_1 models a defender with privileged vulnerability intelligence (EPSS access), whereas R_2 represents a well-informed attacker benefiting from leaked or proprietary exploit data. Comparing these scenarios reveals how unequal access to information shapes convergence speed, strategy stability, and attainable rewards.

The four information configurations were chosen to represent a spectrum of intelligence asymmetry observed in practice. They range from a fully symmetric baseline (R_0) to extreme imbalances where only one side has access to vulnerability intelligence (R_1 or R_2). Scenario R_3 captures mixed or collaborative intelligence environments often seen in modern security ecosystems. This framework provides a structured basis for evaluating how the value of information manifests under adaptive learning dynamics.

3.6 Strategy Pools

To maintain tractable state spaces, each agent chooses actions from a predefined strategy pool (see Table 3.2). Each strategy defines a deterministic resource-allocation pattern across the ten battlefields.

For example, an even strategy distributes resources uniformly across all vendors, while a value-based strategy allocates proportionally to the perceived risk values v_k. A top-3 strategy concentrates effort on the most valuable battlefields, and a random strategy introduces exploratory noise to avoid premature convergence.

The defender's D5 strategy introduces a predictive element: using exponentially decaying memory of past attacks, it forecasts likely attack patterns and reinforces protection where historical trends and high EPSS values overlap. This creates a hybrid rule-based and learning-based mechanism that enriches the adaptive behavior of the defender within the RL loop.

3.7 Simulation Dynamics

The simulation operates on two interconnected timescales.

Inner loop — Reinforcement learning episodes: Within each simulated day, the attacker and defender interact through multiple episodes. Each episode consists of strategy selection, payoff computation, and Q-value updates, allowing gradual adaptation to the current environment.

Outer loop — Daily updates: At the start of each new day, battlefield values v_k are refreshed using the latest aggregated CVE, CVSS, and EPSS data. This reflects the

Table 3.2: Strategy pools of the attacker and defender agents. Each strategy defines a deterministic allocation pattern across the ten battlefields.

ID	Name	Description	Type
		Attacker strategies	
A1	Even allocation	Uniformly distributes resources across all battlefields.	Uniform baseline
A2	Value-proportional	Allocates proportionally to the visible risk values (v_k).	Value-based
A3	Top-3 targeting	Focuses resources on the three highest-valued battlefields.	Focused (top-3)
A4	Random allocation	Randomized distribution introducing exploratory noise.	Exploratory
A5	All-in (Top-1)	Concentrates all resources on the single most valuable battlefield.	Extreme
		Defender strategies	
D1	Even allocation	Uniformly protects all battlefields.	Uniform baseline
D2	Value-proportional	Allocates defense proportionally to perceived risk (v_k).	Value-based
D3	Top-3 defense	Focuses protection on the three highest-valued battlefields.	Focused (top-3)
D4	Random allocation	Randomized protection pattern to reduce predictability.	Exploratory
D5	Predictive counter	Reinforces defense where recent attacks and high EPSS overlap, using exponentially decaying memory.	Adaptive
D6	All-in (Top-1)	Puts all defense on the single most valuable battlefield.	Extreme

continuously evolving vulnerability landscape and introduces non-stationarity into the learning process [7, 23].

The overall process is illustrated in Figure 3.2, which visualizes how the two timescales interact within the simulation. Daily updates modify the environment based on new vulnerability data, while intra-day learning episodes capture the agents' adaptive behavior within that environment.

This bi-level structure couples short-term learning adaptation with long-term environmental change, capturing both tactical evolution and strategic drift in cyber conflicts.

3.8 Algorithmic Formulation

The overall simulation procedure is summarized in Figure 3.3.

This compact pseudocode outlines the bi-level simulation structure, emphasizing daily vulnerability updates and adaptive Q-learning interactions between attacker and defender agents.

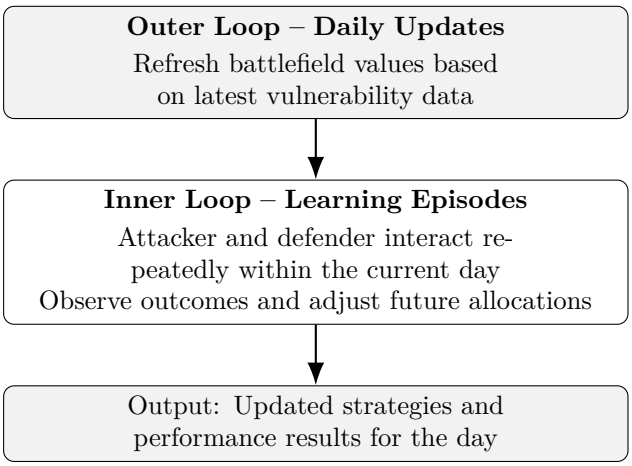

Figure 3.2: Bi-level temporal structure of the simulation (daily updates and intra-day learning episodes).

3.9 Discussion

Beyond its computational behavior, the model also serves as an interpretive framework for analyzing adaptive decision-making in cybersecurity. It demonstrates how defensive learning mechanisms can stabilize or destabilize depending on environmental dynamics, information asymmetry, and feedback quality.

The risk-based formulation $\text{Risk} = \text{CVSS} \times \text{EPSS}$ ensures empirical grounding, while reinforcement learning introduces adaptivity under uncertainty. By combining these elements within the Colonel Blotto framework, the model captures how intelligence availability, bounded rationality, and real-world data collectively influence strategic evolution in cyberspace.

The following chapter defines the experimental configuration and quantitative evaluation metrics used to analyze these dynamics across the four information-asymmetry scenarios.

Algorithm: Dynamic Colonel Blotto Simulation with Reinforcement Learning
Input: CVE, CVSS, EPSS datasets; strategy pools; learning rates α_A, α_D; exploration rates $\varepsilon_A, \varepsilon_D$; scenario parameters R_0–R_3.

1. Initialize $Q_A, Q_D \leftarrow 0$.
2. For each day $t = 1 \dots T$:
   (a) Load vulnerability data and compute $v_k^{(t)} = \text{CVSS}_k^{(t)} \times \text{EPSS}_k^{(t)}$.
   (b) Normalize battlefield values $v_k^{(t)}$.
   (c) For each episode $e = 1 \dots E$:
      i. Select s_A, s_D using ε-greedy on Q_A, Q_D.
      ii. Allocate resources a_k, d_k based on s_A, s_D.
      iii. Determine outcomes with success probability $1 - p_{\text{fail}}$.
      iv. Compute rewards $R^{(\text{att})}$, $R^{(\text{def})}$.
      v. Update:

$$Q_i[s_i] \leftarrow Q_i[s_i] + \alpha_i(R^{(i)} - Q_i[s_i]), \quad i \in \{A, D\}.$$

3. After all episodes: compute mean rewards, win rate, top-3 coverage, dominant (most frequently selected) strategies, and reward ratios.

Figure 3.3: Simplified pseudocode of the reinforcement learning simulation.

Chapter 4

Methodology and Experiments

4.1 Experimental Framework

All experiments were implemented in Python using a modular reinforcement-learning pipeline. The simulation follows the bi-level temporal structure described in Section 3 (Figure 3.2), where multiple learning episodes occur within each simulated day, and daily updates refresh battlefield values using the latest CVSS–EPSS data [5, 7]. This ensures that the environment evolves continuously with real-world vulnerability dynamics.

4.2 Data and Preprocessing

Each simulation uses daily CSV files listing the ten vendors with the highest number of new CVEs for that date. For vendor k, the risk value v_k was computed as defined in Section 3 and normalized using the min–max method. Missing vendors are assigned a zero value to maintain a consistent ten-element vector. All daily vectors are additionally rescaled so that their sum equals ten, ensuring comparability between days while preserving relative differences across vendors. Each normalized snapshot forms one step in the outer time loop of the simulation.

4.3 Simulation Configuration

Within each day, the attacker and defender interact through 500 episodes of tabular Q-learning. This episode count provides sufficient samples for reliable estimation of averaged performance indicators while keeping the overall runtime manageable. Both agents employ an ε-greedy exploration strategy with fixed learning and exploration rates ($\alpha = 0.1$, $\varepsilon = 0.1$), which are common baseline values in the Q-learning literature, ensuring gradual yet responsive adaptation [12]. Resource budgets are asymmetric ($R_A = 20$, $R_D = 40$), reflecting the higher cumulative effort typically required for prevention compared to exploitation. Each attack attempt succeeds with probability $1 - p_{\text{fail}} = 0.7$, introducing moderate stochasticity to approximate uncertainty in real operational environments (e.g., detection errors or heterogeneous system conditions).

The following design choices remain constant throughout all experiments:

- Fixed number of battlefields ($N = 10$) derived from real data.

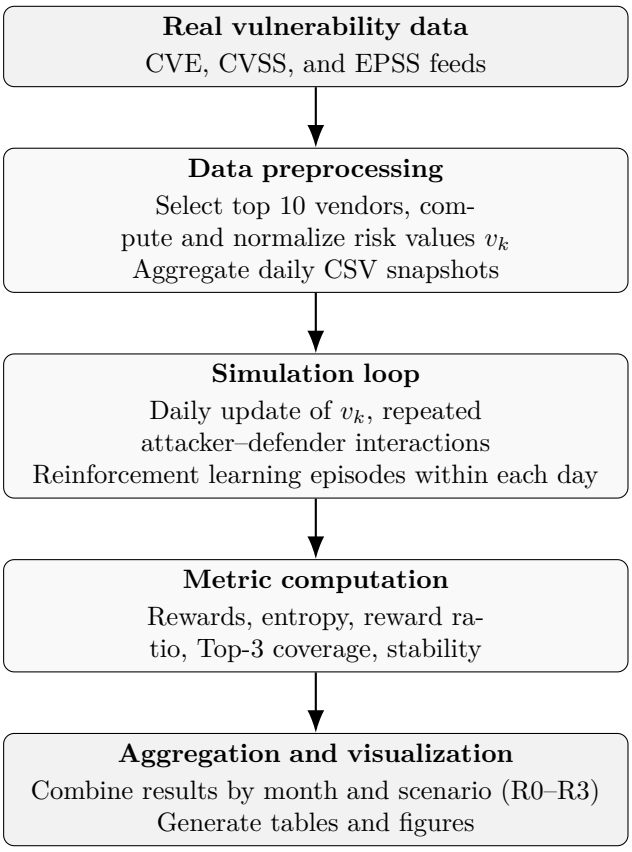

Figure 4.1: Overview of the experimental data-processing and evaluation workflow.

- Daily update of normalized risk values.
- Strategy selection from the predefined pools introduced in Section 3.
- Automatic seed management for reproducibility.

4.4 Batch Execution and Reproducibility

A master script automatically iterated over all combinations of

$$\text{month} \in \{\text{January, May, August}\}, \quad \text{scenario} \in \{R0, R1, R2, R3\}, \quad \text{run} = 1 \dots 10.$$

Each configuration (month × scenario) was repeated ten times to ensure statistical robustness, and results were recorded into unified CSV files containing averaged rewards, stability indicators, and entropy measures. This batch execution structure guarantees experiment isolation while simplifying later aggregation and visualization.

4.5 Evaluation Indicators

To assess agent performance and learning stability, the following quantitative indicators were calculated after each run:

- Average rewards ($\bar{R}_A, \bar{R}_D$) – overall efficiency of attacker and defender.

- Win rate (w_A) – fraction of episodes with at least one successful attack.
- Reward ratio ($\rho = \bar{R}_A/\bar{R}_D$) – balance of achieved payoffs.
- Stabilization episode count (E_{stab}) – number of episodes required for strategy convergence, defined by $\geq 70\%$ dominance and $|\Delta Q| < 10^{-2}$.
- Dominant strategy persistence – fraction of total runtime during which the most frequently selected strategy remains unchanged.
- Shannon entropy of strategy distribution (H_A, H_D) – level of strategic diversity and non-stationarity, computed over 50-episode sliding windows.
- Entropy gap ($\Delta H = H_A - H_D$) – asymmetry in exploration between attacker and defender.
- Top k coverage (C_A^k, C_D^k) – proportion of high-value battlefields successfully controlled or defended at the end of the simulation; $k = 3$ in all runs.

4.6 Measuring Learning Dynamics

In addition to standard performance metrics, the temporal evolution of learning was monitored using the Shannon entropy of the agents' strategy distributions. This indicator quantifies the degree of randomness and exploration in decision-making. High entropy corresponds to diverse or rapidly changing strategies, whereas low entropy indicates temporary stabilization around a dominant allocation pattern. Tracking entropy over time allows the detection of adaptive cycles and mutual reactions between agents [20], highlighting whether their behavior remains dynamic or tends toward stabilization.

4.7 Relating Learning Dynamics to Blotto Equilibrium

In the classical Colonel Blotto game, a Nash equilibrium means that both players use mixed strategies in such a way that neither can improve their payoff by changing their resource distribution alone [14]. Although early analytical solutions existed only for symmetric and continuous variants, recent advances by Loiseau et al. have extended closed-form and computational methods to asymmetric and discrete formulations of the Blotto game [41]. However, this study does not focus on deriving analytical equilibria. Instead, it investigates how reinforcement-learning agents can approximate such equilibria through adaptive, data-driven interaction within a dynamic cyber environment.

In this study, battlefield values change over time based on real vulnerability data. Instead of calculating an exact analytical equilibrium, both players learn and adapt through reinforcement learning. As a result, the Q-learning process reproduces equilibrium-like behavior in simulation, similar to the approach of Noel [20].

Each agent's learned distribution over strategies can be seen as a simulation-based version of a mixed strategy. The distributions remain relatively stable but still fluctuate moderately over time. The repeated use of a few main strategies shows the emergence of a dynamic equilibrium, where both sides keep adapting but neither achieves a lasting advantage. This balance reflects the game-theoretic idea of equal marginal returns across battlefields, reached here through continuous learning instead of formal optimization.

4.8 Summary

In summary, this methodological setup combines real vulnerability data with controlled learning dynamics and reproducible experimentation. It provides a consistent empirical basis for assessing how intelligence asymmetry and environmental dynamics shape attacker–defender adaptation within the proposed game-theoretic framework.

Chapter 5

Results and Discussion

This chapter presents the quantitative results and analysis of the reinforcement learning simulations. All experiments used dynamic battlefields based on the ten vendors with the highest combined CVSS–EPSS risk values. Unless otherwise stated, reported values represent averages across three temporal datasets (January, May, and August) under identical resource settings.

5.1 Aggregate Learning Outcomes

In all simulations, the defender achieved higher cumulative rewards than the attacker. This outcome is partly expected because the defender had a larger resource budget ($R_D = 40$ compared to $R_A = 20$). However, the magnitude of the difference shows that learning and access to information further reinforced the defender's advantage. The mean attacker reward was 1.02 and the defender 8.64, yielding an efficiency ratio

$$\rho = \frac{\bar{R}_A}{\bar{R}_D} \approx 0.12,$$

indicating that only about 12% of the total defended value was compromised. This metric effectively quantifies global protection efficiency, lower values correspond to stronger defensive control.

The attacker's breakthrough rate (the probability that at least one battlefield was breached during an episode) averaged 61.8%, while the success rate of individual attack attempts reached 69%. This difference shows that while local tactical victories were frequent, few translated into strategic impact due to rapid defensive recovery.

Table 5.1: Global performance indicators across all scenarios.

Metric	Mean	Std	Interpretation
Attacker reward $\bar{R}_A$	1.02	0.18	Expected offensive payoff
Defender reward $\bar{R}_D$	8.64	0.31	Cumulative preserved value
Reward ratio ρ	0.12	0.03	Global efficiency of attack
Breakthrough rate	0.618	0.04	Episode-level breach probability
Top-3 coverage (defender)	0.80	0.09	Share of critical assets preserved
Entropy gap $H_A - H_D$	0.27	0.08	Relative strategic diversity

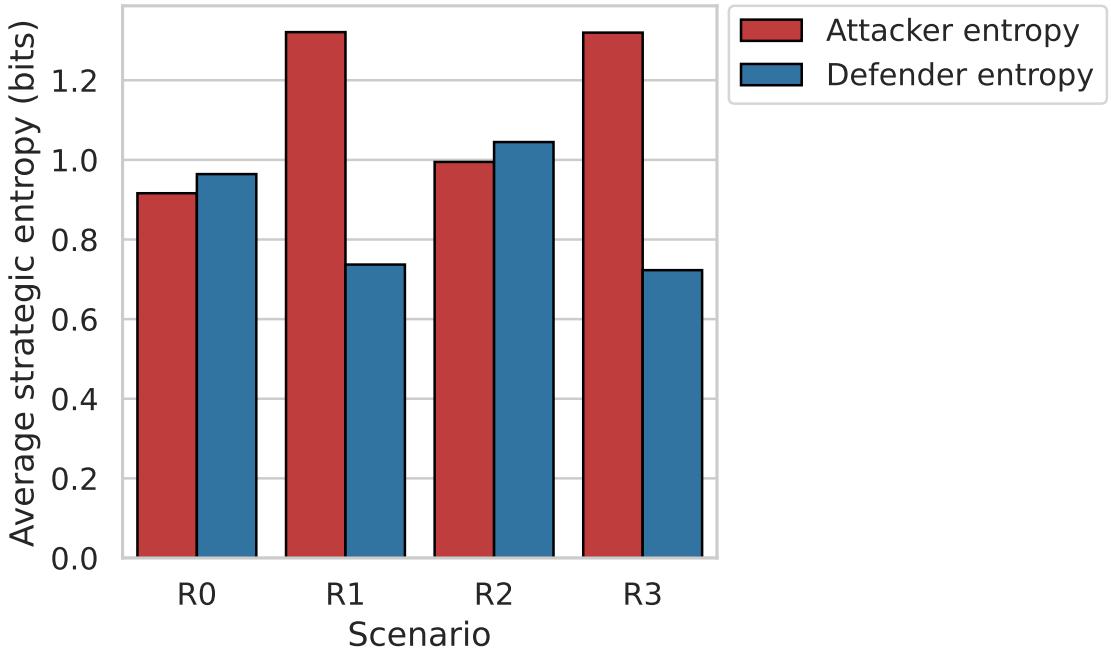

Figure 5.1: Average entropy values of attacker and defender across all scenarios.

The low variance between runs ($< 5\%$) confirms the internal robustness of the reinforcement-learning process. Both agents developed a consistent behavior over time. Exploration introduced small fluctuations, but neither agent became entirely random or fully deterministic.

5.2 Adaptive Dynamics and Strategic Diversity

Entropy-based metrics capture how exploration and adaptation evolved during learning (see Section 4 for the metric definition). Figure 5.1 shows the average entropy trajectories of both agents.

Across all simulations, entropy values remained within a moderate range ($H_A \approx 1.3{-}1.5$, $H_D \approx 1.0{-}1.2$), confirming that both agents continued to explore rather than converging to fixed strategies. The attacker consistently maintained higher entropy ($H_A > H_D$), indicating greater strategic diversity and opportunistic variation, whereas the defender exhibited more focused yet still flexible behavior.

These results suggest a dynamic equilibrium rather than full convergence. Agents stabilized around adaptive policies that balanced exploration and exploitation [20, 36, 35]. Small entropy fluctuations ($\sigma_H < 0.1$) show persistent recalibration rather than instability, demonstrating that effective defense stems from controlled adaptability.

5.3 Strategic Behavior Distributions

Attacker tendencies

The attacker favored concentrated, high-risk strategies. *A3 – Top-3 targeting* dominated (36.7%), followed by *A5 – All-in Top-1* (32.5%), and to a lesser extent *A4 – Random allocation* (18.3%). This outcome reflects the underlying reward structure of the environment. Since only a few high-value battlefields contributed disproportionately to total payoff, the attacker learned that focusing resources on these lucrative targets maximized expected reward. In reinforcement-learning terms, the high variance but occasionally high payoff of concentrated strategies produced stronger positive feedback during exploration, leading to convergence toward top-3 allocations. This mirrors real-world offensive behavior, where limited attack budgets are invested in exploiting the most impactful vulnerabilities rather than maintaining broad but shallow engagement [3, 8].

The partial persistence of the *A4 – Random allocation* strategy illustrates the role of exploration. Even after convergence, stochastic behavior helped the attacker avoid local optima and adapt to dynamic shifts in battlefield values caused by evolving CVE–EPSS distributions. Occasional randomization increased unpredictability, reducing the defender's ability to fully anticipate attack patterns. In practice, this reflects how real adversaries periodically inject noise or decoy operations to mask true intent [38, 18].

The coexistence of *A3*, *A5*, and *A4* therefore represents a multi-modal offensive policy. Measured concentration (*A3*), opportunistic exploitation (*A5*), and controlled randomness (*A4*) together form a balance between exploitation and adaptability. This demonstrates that the attacker's learning process captured not only static reward maximization, but also dynamic resilience against defensive pattern recognition.

Defender tendencies

The defender adopted a more diverse and adaptive portfolio. *D5 – Pattern-reactive defense* appeared most often (41.7%), followed by *D2 – Value-proportional* (29.2%) and *D6 – All-in Top-1* (20.0%). This distribution highlights the defender's ability to combine reactive learning with value-driven prioritization.

At the beginning of training, the defender mostly relied on *D2*, since proportional allocation provided a stable baseline that aligned protection strength with battlefield value. As the simulation progressed, reinforcement signals from repeated attack detections made *D5* more profitable by incorporating short-term memory of recent attack vectors, the model could predict and reinforce likely targets in upcoming episodes. This reactive capability is crucial in adversarial settings where attack patterns exhibit temporal correlation.

The third most frequent strategy, *D6 – All-in Top-1*, emerged in situations where one battlefield's value significantly exceeded the others. In such cases, fully concentrating resources on the single most valuable node was an optimal defensive response, it prevented catastrophic losses even at the cost of partial neglect elsewhere. This is consistent with the structure of EPSS-driven environments, where vulnerability distributions are heavy-tailed and a few assets account for the majority of systemic risk. Hence, the use of *D6* is not random or panicked but a rational risk-minimization tactic under extreme asymmetry, reflecting a "critical asset protection" mindset found in real-world SOC operations [40, 8].

Overall, the defender's strategy portfolio shows a balance between reactive intelligence (D5), baseline prioritization (D2), and extreme focus (D6), demonstrating the agent's

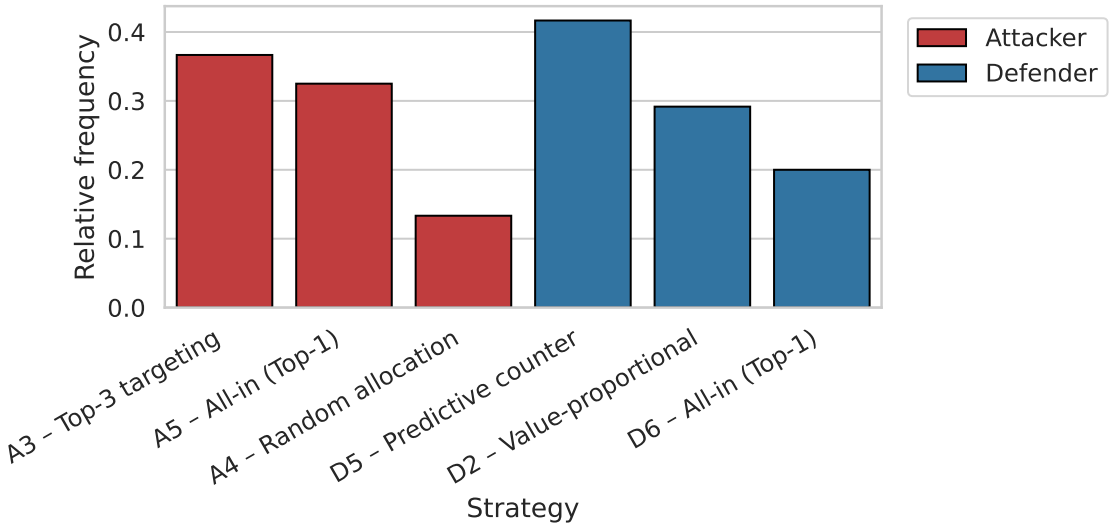

Figure 5.2: Relative frequency of dominant strategies across all runs.

Table 5.2: Scenario-level averages and entropy statistics.

Scenario	$\bar{R}_A$	$\bar{R}_D$	ρ	Breakthrough	H_A	H_D	Top-3 coverage
R0	1.24	8.40	0.15	0.67	0.92	0.96	0.67
R1	0.71	8.98	0.08	0.58	1.32	0.74	0.90
R2	1.41	8.16	0.17	0.67	1.00	1.05	0.74
R3	0.70	9.00	0.08	0.56	1.32	0.72	0.89

capacity to switch between distributed defense and critical-point shielding depending on environmental volatility and observed attack concentration.

The overall strategy distributions of both agents are summarized in Figure 5.2. The figure highlights that while the attacker tended toward concentrated behavior, the defender maintained greater strategic diversity, reflecting asymmetry in objectives between maximizing short-term gain and minimizing cumulative loss through sustained adaptation.

5.4 Information Regime Effects

Each configuration models a distinct visibility structure between the agents. Table 5.2 summarizes the key performance indicators.

Figure 5.3 provides a visual comparison of attacker and defender rewards across the four information regimes, along with their reward ratio (ρ). The figure complements the numerical results in Table 5.2 by showing that the defender consistently maintains a strong advantage ($\bar{R}_D \gg \bar{R}_A$), while the relative efficiency of attacks (ρ) remains below 0.2 in all cases.

R0 – Symmetric baseline

Baseline configuration with equal information access. Rewards ($R_A = 1.24$, $R_D = 8.40$) and near-equal entropy levels ($H_A \approx 0.92$, $H_D \approx 0.96$) indicate balanced mutual adapta-

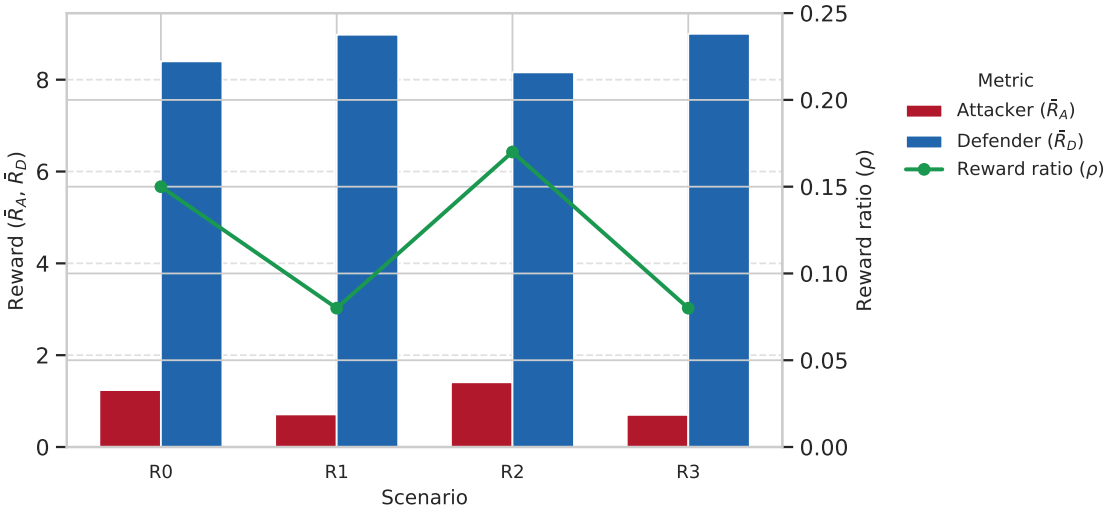

Figure 5.3: Comparison of attacker and defender rewards and reward ratio (ρ) across information regimes.

tion without dominance. This serves as the neutral reference for evaluating the impact of asymmetry. Both agents relied primarily on exploratory learning, reflecting slower adaptation in the absence of privileged intelligence.

R1 – Blue intelligence edge

Defender-side intelligence yielded the strongest protection efficiency. 90% Top-3 coverage, lowest breakthrough probability (58%), and minimum reward ratio ($\rho = 0.08$). Lower entropy ($H_D = 0.74$) reflects more consistent, value-aligned defensive behavior. Access to vulnerability information enabled the defender to allocate resources more effectively and maintain stable protection over time.

R2 – Red intelligence edge

Attacker-side intelligence temporarily increased offensive reward ($R_A = 1.41$), yet overall efficiency ($\rho = 0.17$) remained below the baseline. Higher attacker entropy ($H_A = 1.0$) shows greater behavioral variability rather than sustained improvement. This pattern suggests that information advantages can enhance short-term performance, but long-term stability remains limited as defensive adaptation offsets early gains.

R3 – Hybrid asymmetry

Partial visibility for both agents produced balanced dynamics and the lowest variance across all scenarios ($\sigma_\rho < 0.02$). Reward asymmetry stayed minimal, mirroring real-world environments where information sharing is partial but mutual [19]. This mixed regime represents the most stable interaction pattern, indicating that limited intelligence exchange can sustain adaptive competition without clear long-term dominance.

To test the stability of these outcomes over time, the same experiments were conducted using datasets from January, May, and August. The results confirmed strong temporal robustness. Although the underlying CVE and EPSS distributions varied seasonally, the

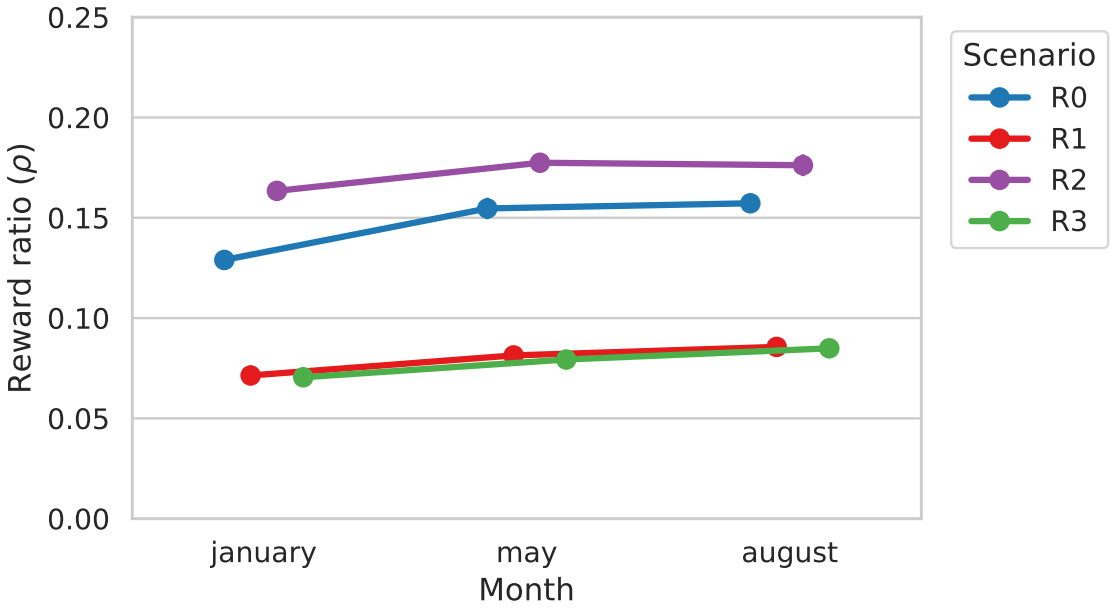

Figure 5.4: Cross-month consistency of scenario efficiency based on reward ratio (ρ) trends.

relative ranking of scenarios remained consistent, with $R2$ and $R0$ showing higher reward ratios and $R1$ and $R3$ maintaining the lowest values. This indicates that attacker-side intelligence (R2) consistently increased offensive efficiency, while defender-informed configurations (R1, R3) retained superior defensive control.

As shown in Figure 5.4, the reward ratio values remained nearly constant across months, demonstrating that the learned strategies generalized effectively beyond a single period of vulnerability data.

5.5 Top-3 Coverage and Value Retention

The Top-3 coverage metric connects simulation outcomes with practical defense relevance. As shown in Figure 5.5, defenders consistently retained the majority of high-value assets. On average, they preserved 80% of the cumulative value of the three most critical vendors, nearly quadruple the attacker's share. In $R1$, coverage reached 90%, implying that intelligence-driven prioritization enables near-total protection of top-value assets.

The marginal gains beyond the top-3 decrease sharply, indicating that a small number of assets account for the majority of the overall exposure. The strong correlation between vendor value and defender allocation ($r > 0.85$) further supports that the RL-based defender learned value-aligned prioritization.

5.6 Interpretation and Summary of Findings

The quantitative results reveal several overarching patterns and implications. First, the presence of information advantage substantially amplifies defensive efficiency. In defender-informed scenarios (R1 and R3), breach probabilities decreased by up to 15 percentage

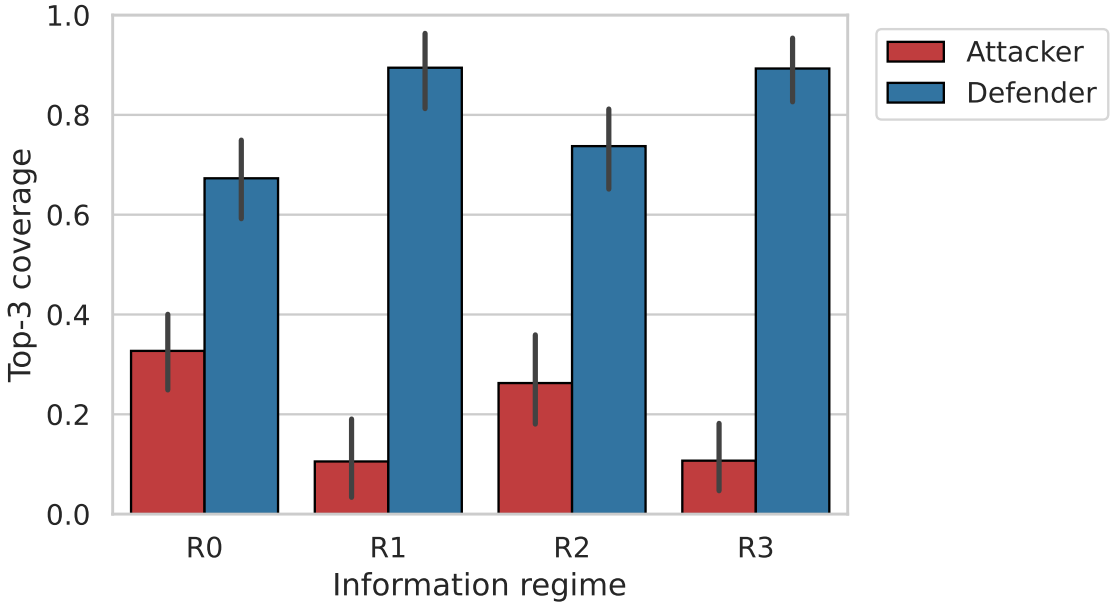

Figure 5.5: Top-3 coverage across information regimes.

points compared to configurations without intelligence support. This confirms that access to vulnerability intelligence enables more precise and stable resource allocation over time.

Second, instead of converging to fixed strategies, both agents co-evolved within a dynamic equilibrium. This oscillatory regime illustrates that adaptive learning, rather than static optimization, governs sustained cyber competition. Neither side achieved permanent dominance, instead, both continually recalibrated their strategies in response to the other's behavior and environmental shifts.

Entropy measures further highlighted the role of adaptability as a resilience indicator. Sustained mid-level entropy values reflected a balance between exploration and exploitation. Defenders avoided rigidity while attackers maintained unpredictability. Together, Top-3 coverage (C_D) and defender entropy (H_D) serve as complementary metrics of quantitative resilience, capturing both the protection of critical assets and the adaptive flexibility of defensive posture.

From a practical perspective, the results suggest that reinforcement-learning architectures of this kind could underpin automated threat-prioritization systems or serve as red–blue simulation frameworks for evaluating adaptive strategies in cyber defense. The defender's preserved value exceeded the attacker's gain by nearly an order of magnitude, and intelligence asymmetry (R1 vs. R2) explained most performance variance. Defender-side intelligence improved efficiency by roughly +0.16 in reward ratio. No convergence to static equilibria was observed, and strategy oscillations persisted throughout all episodes. Top-3 coverage averaged 80%, peaking at 90% in R1, while moderate entropy confirmed sustained adaptability rather than random fluctuation. These results also remained robust across multiple temporal datasets, confirming the generalizability of the learned behaviors.

In summary, the results demonstrate that in cyber environments driven by real-time vulnerability intelligence, optimal defense does not emerge from static equilibrium but from ongoing, informed adaptation. The dynamic interplay between learning agents captures the essence of modern cybersecurity, representing resilience through continuous evolution rather than fixed strategies.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

This research presented a dynamic Colonel Blotto framework that models cyber conflicts as adaptive resource-allocation games over real-world vulnerability data. By integrating daily CVSS–EPSS updates into a reinforcement learning simulation, the model captured how bounded-rational attacker and defender agents evolve their strategies under varying levels of intelligence asymmetry.

The results demonstrate several key insights. First, the dynamic nature of the battlefield values prevented static equilibria from emerging, forcing both agents to adapt continuously. The defender's advantage in information access (R_1 scenario) consistently produced more coherent adaptation patterns and higher average payoffs, while the attacker's information edge (R_2) yielded more volatile learning curves and slower convergence. When both agents had partial visibility (R_3), the game exhibited oscillatory dynamics and increased entropy, suggesting that mutual adaptation can maintain a state of dynamic equilibrium rather than convergence to a fixed strategy.

Empirical reinforcement-learning results supported these conceptual findings. Average reward ratios indicated that the defender maintained a positive balance in most simulations, especially when predictive counter-strategies (D5) were available. Entropy trends revealed alternating phases of exploration and consolidation, showing that neither agent fully dominated the strategic landscape over extended periods. Top-3 coverage results further confirmed that defenders with intelligence access protected a larger share of high-risk vendors, emphasizing the operational value of vulnerability intelligence in real-time decision-making.

In summary, the study contributes a data-driven, adaptive, and reproducible approach to modeling attacker–defender interactions. It bridges the gap between theoretical game formulations and empirical cyber-risk dynamics, illustrating how information asymmetry, reinforcement learning, and vulnerability dynamics jointly shape cyber-strategic evolution.

These findings directly address the initial research objectives. The integration of real CVSS–EPSS dynamics (RO1) successfully produced a temporally evolving, non-stationary environment representative of actual vulnerability trends. Reinforcement-learning agents (RO2) replicated adaptive attacker–defender interactions, achieving dynamic equilibrium rather than static convergence. The information-asymmetry experiments (RO3) quantified measurable advantages of defender-side intelligence and demonstrated its stabilizing effect on reward variance. Finally, entropy-based metrics (RO4) proved effective in iden-

tifying structured adaptation and resilience under uncertainty, validating their relevance as indicators of strategic flexibility in complex cyber environments.

6.2 Limitations

While the model captures essential aspects of cyber conflict dynamics, several simplifications limit its scope. First, the resource-allocation space was discretized into a fixed set of predefined strategies, which constrained exploration and limited the emergence of novel hybrid behaviors. Second, the probabilistic payoff mechanism relied solely on an attacker-side success parameter (p_{fail}), without modeling heterogeneous exploit reliability or defensive response latency. Third, the asymmetry between attacker and defender resources was kept constant ($R_A = 20$, $R_D = 40$) across all simulations. In reality, this balance can vary widely. A well-funded state-level attacker may face a small private organization with limited defense capacity, while in other settings a strong corporate security team may defend against opportunistic adversaries. Allowing the resource ratio to change dynamically or across scenarios could provide deeper insight into how relative power and capability asymmetries affect adaptation and equilibrium outcomes. Fourth, the use of daily vendor-level CVSS–EPSS data aggregates complex vulnerability ecosystems into single scalar values, neglecting inter-vendor dependencies and cross-domain effects. Finally, the agents operated in a zero-communication environment without modeling deception, misinformation, or shared intelligence, all of which are crucial in real-world cyber operations.

Despite these limitations, the framework provides a robust foundation for studying adaptive decision-making under realistic data conditions and partial observability.

6.3 Future Work

Future research can extend the current model along several dimensions.

- **Continuous strategy spaces:** Replace the discrete strategy pools with parameterized policies or neural approximators, allowing deep Q-learning or actor–critic methods to discover novel allocation patterns.
- **Hierarchical temporal structure:** Introduce multi-day or weekly planning layers where agents learn not only per-day allocations but also long-term adaptation policies over changing vulnerability landscapes.
- **Richer intelligence modeling:** Expand the information-asymmetry regimes (R_0–R_3) to include misinformation, delayed reporting, or probabilistic intelligence leaks, thereby simulating deception and uncertainty propagation.
- **Graph-based battlefield representation:** Replace independent vendors with a connected graph reflecting software dependencies or supply-chain relationships, enabling the analysis of cascading effects and correlated risks.
- **Two-stage attack modeling:** Introduce a two-phase attack process with an initial exploration step to probe defenses at a cost, followed by exploitation based on the gathered intelligence. This extension would capture the trade-off between reconnaissance and direct attack in adversarial strategy design.

- **Real-time data integration:** Link the simulation directly to live EPSS and NVD feeds to evaluate daily model adaptation on operational cyber-threat intelligence streams.
- **Human-in-the-loop evaluation:** Combine automated agents with human decision feedback to explore mixed-initiative defense strategies and cognitive biases in cyber-resource allocation.

By pursuing these directions, the dynamic Colonel Blotto framework can evolve into a comprehensive experimental platform for testing adaptive cyber-defense policies, risk-aware learning algorithms, and information-sharing mechanisms in complex digital ecosystems.

All data, scripts, and simulation code used in this study are publicly available in the following GitHub repository: `https://github.com/nazalan/cyber-blotto-rl`.

Bibliography

[1] John Arquilla and David Ronfeldt. Cyberwar is coming! *Comparative Strategy*, 12(2):141–165, 1993.

[2] Thomas Rid. *Cyber War Will Not Take Place*. Oxford University Press, 2013.

[3] Rainer Böhme and Márk Félegyházi. Optimal information security investment with penetration testing. *IFIP WG 11.10 International Conference on Critical Infrastructure Protection*, 2009.

[4] Rainer Böhme and Tyler Moore. The iterated weakest link: A model of adaptive security investment. In *Workshop on the Economics of Information Security (WEIS)*, 2009.

[5] MITRE Corporation. Common vulnerabilities and exposures (cve) overview. `https://cve.mitre.org/`, 2024. Accessed: 2025-10-28.

[6] FIRST Organization. *Common Vulnerability Scoring System v2.0 Guide*. Forum of Incident Response and Security Teams (FIRST), 2007. ITU-T Recommendation X.1521, `https://www.first.org/cvss/v2/guide`.

[7] Jay Jacobs, Sasha Romanosky, Benjamin Edwards, Idris Adjerid, and Michael Roytman. Exploit prediction scoring system (epss). *Digital Threats: Research and Practice*, 2(3):20:1–20:17, 2021.

[8] Kjell Hausken, Jonathan W. Welburn, and Jun Zhuang. A review of attacker–defender games and cyber security. *Games*, 15(4):28, 2024.

[9] Anibal Sanjab and Walid Saad. On bounded rationality in cyber-physical systems security: Game-theoretic analysis with application to smart grid protection. *arXiv preprint arXiv:1610.02110*, 2016. Submitted to IEEE Transactions on Smart Grid.

[10] Ross Anderson. Why information security is hard – an economic perspective. In *Proceedings of the 17th Annual Computer Security Applications Conference (ACSAC)*, pages 358–365. IEEE, 2001.

[11] Ross Anderson and Tyler Moore. The economics of information security. *Science*, 314(5799):610–613, 2006.

[12] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2nd edition, 2018.

[13] Christopher J. C. H. Watkins and Peter Dayan. Q–learning. *Machine Learning*, 8(3–4):279–292, 1992.

[14] Brian Roberson. The colonel blotto game. *Economic Theory*, 29:1–24, 2006.

[15] Pern Hui Chia and John Chuang. Colonel blotto in the phishing war. In *Proc. of the Workshop on the Economics of Information Security (WEIS)*, 2008.

[16] Aidin Ferdowsi and Walid Saad. A colonel blotto game for interdependent security in cyber–physical systems. In *Proceedings of GameSec 2017*, 2017.

[17] G. J. Kim and Hyunsoo Park. Sequentially variant colonel blotto game with one–sided and incomplete information. *IEEE Transactions on Games*, 2024.

[18] Tyler Moore, Allan Friedman, and Ariel D. Procaccia. Would a 'cyber warrior' protect us? exploring trade-offs between attack and defense of information systems. In *Workshop on the Economics of Information Security (WEIS)*, 2010.

[19] Abhishek Gupta, Tamer Başar, and Galina A. Schwartz. A three-stage colonel blotto game: When to provide more information to an adversary. In *Decision and GameTheory for Security – 5th International Conference, GameSec 2014, Proceedings (LNCS 8840)*, 2014.

[20] Joseph Christian G. Noel. Reinforcement learning agents in colonel blotto. *arXiv preprint arXiv:2204.02785*, 2022.

[21] Amrin Maria Khan Adawadkar and Nilima Kulkarni. Cyber-security and reinforcement learning – a brief survey. *Engineering Applications of Artificial Intelligence*, 114, 2022.

[22] Soheil Behnezhad, Mohammad Taghi Hajiaghayi, and Saeed Seddighin. The auditing game: Mechanism design for network security audits. In *Proceedings of the 2018 International Conference on Game Theory for Networks (GameNets)*, 2018.

[23] Luca Allodi and Fabio Massacci. Comparing vulnerability severity and exploitability metrics. *ACM Transactions on Information and System Security (TISSEC)*, 17(1):1–20, 2014.

[24] Mehran Bozorgi, Lawrence K. Saul, Stefan Savage, and Geoffrey M. Voelker. Beyond heuristics: Learning to classify vulnerabilities and predict exploits. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 105–114, 2010.

[25] MITRE Corporation. Common vulnerabilities and exposures (cve) repository. `https://cve.mitre.org/`, 2024. Accessed: 2025-11-01.

[26] National Institute of Standards and Technology (NIST). National vulnerability database (nvd). `https://nvd.nist.gov/`, 2024. Accessed: 2025-11-01.

[27] OffSec. Exploit database (exploit-db). `https://www.exploit-db.com/`, 2024. Accessed: 2025-11-01.

[28] Rapid7. Metasploit framework exploit database. `https://www.metasploit.com/`, 2024. Accessed: 2025-11-01.

[29] D2 Security. D2 elliot exploitation framework. `https://d2sec.com/`, 2024. Accessed: 2025-11-01.

[30] Immunity Inc. Canvas exploitation framework. `https://www.immunityinc.com/products/canvas/`, 2024. Accessed: 2025-11-01.

[31] Proofpoint Inc. Proofpoint threat intelligence and telemetry feeds. https://www.proofpoint.com/us/threat-insight, 2024. Accessed: 2025-11-01.

[32] Fortinet Inc. Fortiguard threat intelligence and research. https://www.fortiguard.com/, 2024. Accessed: 2025-11-01.

[33] AlienVault (AT&T Cybersecurity). Alienvault open threat exchange (otx). https://otx.alienvault.com/, 2024. Accessed: 2025-11-01.

[34] Nicolo Cesa–Bianchi, Claudio Gentile, and Gabor Lugosi. Boltzmann exploration done right. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[35] Yue Li, Kaiqing Zhang, and Tamer Basar. Adaptive ε–greedy strategies for reinforcement learning in dynamic games. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[36] Zijian An and Lifeng Zhou. Reinforcement learning for game-theoretic resource allocation on graphs. *arXiv preprint arXiv:2505.06319*, 2025. Forthcoming.

[37] Aron Laszka, Sadegh Farhang, and Jens Grossklags. On the economics of ransomware. In *Proceedings of the Conference on Decision and Game Theory for Security (GameSec)*, 2017.

[38] Tongxin Yin, Armin Sarabi, and Mingyan Liu. Deterrence, backup, or insurance: Game-theoretic modeling of ransomware. *IEEE Transactions on Information Forensics and Security*, 2020.

[39] Róbert Mucsi. Analyzing cyber conflicts: Battlegrounds, resource asymmetry and real-world data. Master's thesis, Budapest University of Technology and Economics, 2023.

[40] European Union. Directive (eu) 2022/2555 of the european parliament and of the council on measures for a high common level of cybersecurity across the union (nis2 directive). https://eur-lex.europa.eu/eli/dir/2022/2555, 2022. Accessed: 2025-10-29.

[41] Dong Quan Vu, Patrick Loiseau, and Alonso Silva. A simple and efficient algorithm to compute ε-equilibria of discrete colonel blotto games. In *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018)*, pages 2115–2118, 2018.