TALLINN UNIVERSITY OF TECHNOLOGY
School of Information Technologies

Kristo Tammsoo 232672IVCM

EVALUATING BLUE TEAM TTPS AGAINST CYBER THREATS ON THE EXAMPLE OF LOCKED SHIELDS 2024

Master's Thesis

Supervisor: Rain Ottis
PhD

Co-supervisor: Bernt Åkesson
D.Sc. (Tech.)

Tallinn 2026

TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond

Kristo Tammsoo 232672IVCM

SINISE MEESKONNA TTP-DE HINDAMINE KÜBEROHTUDE VASTU LOCKED SHIELDS 2024 NÄITEL

Magistritöö

Juhendaja: Rain Ottis
PhD

Kaasjuhendaja: Bernt Åkesson
D.Sc. (Tech.)

Tallinn 2026

Author's Declaration of Originality

I hereby certify that I am the sole author of this thesis. All the used materials, references to the literature and the work of others have been referred to. This thesis has not been presented for examination anywhere else.

Author: Kristo Tammsoo

January 9, 2026

Abstract

Cybersecurity exercises like NATO's Locked Shields provide platforms for testing and refining defensive tactics, techniques, and procedures (TTPs) against sophisticated cyber threats. This thesis systematically evaluates the effectiveness of Blue Team TTPs used during the Locked Shields 2024 exercise. A mixed-methods approach was adopted, integrating qualitative data obtained through semi-structured interviews with quantitative performance metrics derived from the official Final Exercise Report (FER). Interviews with key Blue Team personnel allowed for in-depth exploration of strategic decisions, organizational structures, automation integration, and communication processes. These qualitative insights were subsequently correlated with quantitative exercise outcomes, highlighting effective practices and identifying areas requiring improvement. The study finds that successful Blue Teams effectively leveraged automation, maintained operational flexibility through structured internal coordination, and balanced preventive, detective, and responsive capabilities. Challenges such as communication bottlenecks, insufficient pre-exercise preparations, and limitations in automation were also identified. This research contributes a reproducible framework for systematically evaluating Blue Team strategies and provides insights to enhance cybersecurity readiness in both exercise and real-world contexts.

The thesis is written in English and is 83 pages long, including 7 chapters, 6 figures and 7 tables.

Annotatsioon

Sinise meeskonna TTP-de hindamine küberohtude vastu Locked Shields 2024 näitel

Küberkaitseõppused õppused, nagu NATO Locked Shields, pakuvad keskkondi keerukate küberohtude vastu suunatud kaitsetaktika, -tehnika ja -protseduuride (TTP) testimiseks ja täiustamiseks. Käesolevas magistritöös hinnatakse süstemaatiliselt õppuse Locked Shields 2024 käigus kasutatud sinise meeskonna TTPde tõhusust. Kasutati segameetodilist lähenemisviisi, ühendades poolstruktureeritud intervjuude kaudu saadud kvalitatiivsed andmed ametlikku õppuse lõpparuandest (FER) saadud kvantitatiivsete tulemuslikkuse näitajatega. Intervjuud sinise rühma võtmeisikutega võimaldasid põhjalikult uurida strateegilisi otsuseid, organisatsioonilisi struktuure, automaatika integreerimist ja kommunikatsiooniprotsesse. Neid kvalitatiivseid teadmisi seostati seejärel kvantitatiivsete õppuse tulemustega, tuues esile tõhusad tavad ja parandamist vajavad valdkonnad. Uuringus leitakse, et edukad sinised meeskonnad kasutasid tõhusalt ära automatiseerimist, säilitasid operatiivset paindlikkust struktureeritud sisemise koordineerimise kaudu ning tasakaalustasid ennetus-, avastamis- ja reageerimisvõimet. Samuti tuvastati selliseid probleeme nagu kommunikatsioonipuudujäägid, ebapiisav õppuseelne ettevalmistus ja automatiseerimise piirangud. Käesolev magistritöö aitab luua korratava raamistiku siniste tiimide strateegiate süstemaatiliseks hindamiseks, andes nii õppustel kui ka päriselus rakendatavaid teadmisi küberkaitse valmisoleku suurendamiseks.

Lõputöö on kirjutatud inglise keeles ning sisaldab teksti 83 leheküljel, 7 peatükki, 6 joonist, 7 tabelit.

List of Abbreviations and Terms

BCS	Baltic Cyber Shield
BT	Blue Team
C2	Command and Control
CCDC	Collegiate Cyber Defence Competitions
CCDCOE	Cooperative Cyber Defence Centre of Excellence
CDN	Content Delivery Network
CTF	Capture the Flag
DMZ	Demilitarized Zone
EDR	Endpoint Detection and Response
FER	Final Exercise Report
GT	Green Team
GPO	Group Policy Object
ICS	Industrial Control System
IDS	Intrusion Detection System
IPS	Intrusion Prevention System
LS	Locked Shields
LLM	Large Language Model
NATO	North Atlantic Treaty Organization
OAT	Observational Assessment of Teamwork
RCE	Remote Code Execution
RDP	Remote Desktop Protocol
RT	Red Team
SIEM	Security Information and Event Management
SINET	Simulated Internet
SOC	Security Operations Centre
TTP	Tactics, Techniques, and Procedures
UST	User Simulation Team
VM	Virtual Machine
WAF	Web Application Firewall
XSS	Cross-site scripting

Table of Contents

List of Figures

List of Tables

1. Introduction

Cybersecurity is a priority for nations and organizations, particularly as cyber threats grow in complexity and sophistication [1]. As a result, advanced exercises designed to simulate real-world cyberattacks help develop and test defensive strategies. Locked Shields, an annual exercise organized by the NATO Cooperative Cyber Defence Centre of Excellence (CCDCOE), provides an environment for this purpose. Since 2010, Locked Shields has offered cybersecurity professionals the opportunity to hone their skills in defending national IT systems and critical infrastructure against live, simulated attacks [2]. Through such exercises, defenders are better equipped to respond to genuine threats, strengthening cyber defences across industries and national borders.

The scale and realism of Locked Shields elevate its value as a training and evaluation platform. For example, the 2024 edition brought together some 4,000 experts from more than 40 nations to defend virtualised critical-infrastructure systems [3]. The nature of a live-fire exercise demands that Blue Teams not only apply technical controls, but also manage crisis communications, legal aspects, and strategic decision-making in real time. As such, it reflects the multifaceted nature of modern cyber defence, where technical skill alone is insufficient without coordination, situational awareness and organisational resilience.

This thesis examines the tactics, techniques, and procedures (TTPs) used by Blue Teams during Locked Shields 2024. Understanding the effectiveness of these defences helps to fortify current strategies and also offers insights into areas where improvements are needed.

Through interviews with key Blue Team members, an analysis of their chosen TTPs will be created. The findings aim to validate the current TTPs used in critical infrastructure defence and highlight those most effective in crafting a robust cyber defence framework.

1.1 Problem statement

The increasing complexity and frequency of cyberattacks pose significant challenges to organizations worldwide, highlighting the need for effective defensive strategies. Blue Teams, responsible for safeguarding systems during live-fire exercises use different TTPs to detect, mitigate, and respond to advanced threats. However, there is limited research

evaluating the effectiveness of these TTPs in live-fire cybersecurity exercise scenarios.

The absence of a structured analysis of Blue Team TTPs not only limits academic understanding but also affects the practical development of robust defence mechanisms. By addressing this gap organizational readiness against evolving cyber threats can be improved. This research seeks to systematically evaluate Blue Team TTPs during Locked Shields, by conducting semi-structured interviews with key Blue Team members, to identify effective practices, highlight inefficiencies, and provide actionable insights for improving real-world and exercise-related cyber defence strategies.

1.2 Assumptions and Research Questions

The assumptions and research questions for this study are designed to explore the effectiveness of Blue Team TTPs during the Locked Shields 2024 exercise. They aim to establish a systematic understanding of the relationship between specific TTPs and their outcomes, identify strategies that are consistently successful, and discover areas for improvement in cyber defence practices. The following assumptions and research questions guide this investigation:

1.2.1 Assumptions

The underlying principles of this thesis's analytical framework are based on the following assumptions. They serve as the basis for evaluating Blue Team TTPs within the context of the Locked Shields 2024 exercise:

- **A1:** The effectiveness of Blue Team TTPs in mitigating cyber threats during Locked Shields 2024 can be systematically evaluated and correlated with specific defence outcomes.
- **A2:** Certain TTPs are consistently more effective in mitigating specific types of cyber threats.
- **A3:** The evaluation of TTPs will reveal gaps or inefficiencies that can inform recommendations for improved defence strategies in real-world and live-fire exercises.

1.2.2 Research Questions

To guide the investigation, the following research questions have been formulated. These questions aim to evaluate the applicability and efficacy of Blue Team practices in the context of Locked Shields 2024:

- **RQ1:** How can the effectiveness of Blue Team TTPs during the Locked Shields 2024 exercise be systematically evaluated and correlated with specific defence outcomes?
- **RQ2:** What strategies can be recommended to improve Blue Team coordination and decision-making during live-fire exercises, based on the evaluation of TTP performance?

1.3 Research Goals

The primary goal of this study is to systematically evaluate the effectiveness of Blue Team TTPs during the Locked Shields 2024 exercise, providing a reproducible framework for analyzing defensive strategies in live-fire cybersecurity scenarios. In addition, the study seeks to identify the gaps or inefficiencies in current TTPs, offering recommendations to improve the coordination, decision making, and allocation of resources in Blue Teams. The findings are intended to be adaptable, contributing not only to the design of future cybersecurity exercises, but also to improving real-world organizational readiness against evolving cyber threats.

1.4 Outline of the thesis

This thesis is structured into chapters, each contributing to a comprehensive analysis of Blue Team TTPs within the context of the Locked Shields 2024 exercise. The following provides a brief overview of the content and purpose of each chapter:

- **Literature Review**
  Investigates prior work on Blue Team operations, cyber defence frameworks, and live-fire cyber exercises, providing the theoretical and contextual foundation for the study.
- **Methodology**
  Describes the research design, including data collection and analysis methods. It explains the rationale for using interviews with Blue Team members and outlines the approach to evaluating TTP effectiveness.
- **Results**
  Presents empirical findings based on the collected data. It summarizes the key TTPs observed and links them to relevant defence outcomes during the exercise.
- **Evaluation of Blue Team TTPs**
  Analyzes the results in relation to the research questions. It evaluates the strengths, weaknesses, and effectiveness of specific TTPs, identifying patterns and opportunities for improvement.

- **Discussion**
  Interprets the findings in a broader context, discussing implications for real-world cyber defence and reflecting on the limitations and potential generalizability of the study.
- **Summary**
  Concludes the thesis by summarizing key insights and contributions. It also proposes directions for future research and offers practical recommendations for enhancing Blue Team practices.

2. Literature Review

There is a research gap in the systematic evaluation of Blue Team TTPs within the context of live-fire exercises like Locked Shields. Although general defensive practices have been studied, there is a lack of in-depth analysis correlating specific TTPs to their effectiveness against particular types of cyber threats. The identified research gap is directly aligned with the research problem in this thesis, which seeks to address the lack of systematic evaluation of Blue Team TTPs.

2.1 Cybersecurity Exercises

Cybersecurity exercises have proven to be a critical component in converting theoretical defence strategies into practical operational readiness. For example, the study by Smeets [4] highlights several key points of the Locked Shields live-fire exercise, as to what value cybersecurity exercises bring:

- Live-fire exercises expose system vulnerabilities and test incident response capabilities under pressure.
- Such simulations enhance the development of advanced technical skills and strategic coordination among defence teams.
- Challenges remain in training for complex technical scenarios in a virtual environment and in managing long-term campaign planning.

In a study by Kern et al. [5] a novel approach is presented to enhancing cybersecurity tooling, specifically Intrusion Detection Systems (IDS), through Capture-the-Flag (CTF) exercises. Similarly, live-fire cybersecurity exercises can expand this approach by testing how effectively Blue Teams can correlate logs and transform information flows into actionable defence decisions under real-time pressure.

Švábenský et al. [6] provide a comprehensive review of tabletop exercises, emphasizing their significance in preparing cybersecurity personnel for effective incident response and resolution. Their study demonstrates that these exercises systematically uncover procedural inefficiencies and process gaps, directly enhancing organizational preparedness. Additionally, the insights derived from tabletop simulations serve as valuable inputs for refining cybersecurity tools, processes, and strategic decision-making capabilities.

Mäses argues that many cybersecurity exercises remain underutilised because they often prioritise point-based outcomes and competition-oriented scoring systems, while overlooking the deeper behaviours, competencies, and collaborative processes that drive defensive success [7]. This critique is particularly relevant in the context of Locked Shields, where Blue Teams must simultaneously manage technical defences, maintain services, communicate with other teams and exercise control elements, and respond to legal or strategic injects. Maennel further observes that evidence-based assessment in exercises is frequently limited, as many after-action evaluations rely on subjective interpretation rather than systematically collected digital traces or behavioural analytics, resulting in a gap between exercise design and meaningful measurement of learning or defensive capability [8]. These findings underscore the importance of structured evaluation frameworks—such as those used in this thesis—that combine qualitative insights with quantitative metrics.

Parsons et al. demonstrate that human factors—particularly information security awareness and behaviour—remain critical determinants of defensive performance, with human error implicated in the vast majority of breaches [9]. Ernits et al. highlight that realistic, data-driven cyber exercises enable measurable skill development by linking technical event logs to higher-level competencies, thereby offering a structured basis for evaluating defensive proficiency in live-fire scenarios [10]. Similarly, a SCADA security survey underscores how the convergence of operational technology with corporate networks increases systemic vulnerability to coordinated cyber operations, reinforcing the need for rigorous, scenario-based defence training [11]. This is directly relevant to Locked Shields, where similar event-log mechanisms are used in the scoring of automated Red Team attacks to determine whether specific exploits against Blue Team services were successful. Broader definitional work on cybersecurity further argues that cyber defence must be understood as a multi-layered construct combining technological safeguards, human behaviour, and organizational processes, rather than solely a technical domain, necessitating evaluation methodologies that capture this full scope [12].

Overall, the findings suggest that cybersecurity exercises are essential not only for developing and refining the tactical skills and technical proficiency of security professionals but also for enhancing organizational readiness, strategic decision-making capabilities, and resilience against increasingly sophisticated cyber threats.

2.1.1 Locked Shields Exercise

The following description of the exercise and its objectives has been derived from the Final Exercise Report (FER) [13]. Locked Shields (LS) is the name of the world's most advanced live-fire cyber defence exercise [13]. It provides a real-world-like environment

for international cybersecurity specialists to enhance their skills in safeguarding national IT infrastructures and essential assets. The exercise prioritizes realistic scenarios and advanced technological integration, encompassing every aspect of a major cyber incident—from strategic planning and digital forensics to legal considerations and strategic communications. Designed specifically to test participants' responses to complex cyberattacks under realistic pressure conditions, the exercise ensures preparedness against diverse cyber threats. Moreover, the international collaboration fostered throughout the exercise not only advances individual technical proficiency but also reinforces global cyber resilience through the exchange of expertise and effective practices.

Objectives and Format

The main goal of Locked Shields is to provide a platform for enhancing cyber defence skills through realistic crisis simulations. It serves as a testing environment for developing, evaluating, and refining strategic solutions. Locked Shields encourages close collaboration and information sharing among technical specialists and strategic leaders.

The exercise has a competitive yet collaborative format, featuring realistic attack scenarios on a wide array of different classes of endpoints. Participants face challenges requiring them to integrate technical expertise, strategic decision-making, and effective communication. Each participating team is tasked with defending simulated national infrastructure against sophisticated cyberattacks, with performance measured across technical and soft domains, including but not limited to incident handling, digital forensics, and strategic communication effectiveness. Through this format, Locked Shields enhances the participants' ability to respond to threats as an unified team dynamically under pressure, fostering both individual skills development and collective resilience.

Exercise Objectives

The exercise objectives as laid out in the FER are as follows [13]:

- Train teams of cyber professionals to collaborate with each other, to detect and mitigate large-scale cyber-attacks, and to handle security incidents according to the training objectives.
- Provide training for a non-technical audience in legal, strategic communications, and strategic decision-making processes and procedures to respond to various cyber related activities.
- Learn from the activities of the Blue and Red Teams. In cases of similar real-world scenarios, determine which defensive tactics, tools, and procedures work best and

what to expect from attackers.

- Strengthen the international security community by building trust networks, as well as sharing information and experiences.
- Provide training in cyber situational awareness.

2.2 Evaluation of Cyber Defence Tactics and Strategies (TTPs)

This section provides a detailed review of existing methodologies, frameworks, and metrics used to evaluate Blue Team TTPs. By examining the current approaches, the objective is to highlight how effectiveness in cyber defence has been measured and assessed in previous studies, identifying both strengths and limitations in existing evaluation practices. Additionally, this section aims to discuss empirical research focused explicitly on Blue Team performance, setting the foundation for understanding the existing landscape and finding gaps within.

2.2.1 Methodologies and Frameworks for TTP Evaluation

Champion et al. [14] show that the effectiveness of defensive TTPs is significantly influenced by the functional specialization and role composition within Blue Teams. Their research indicates that clearly defined, task-relevant roles—whether shared among team members or uniquely assigned—directly shape how TTPs are planned, executed, and adapted during cyber defence operations. Effective role specialization ensures efficient coordination, enhances situational awareness, and facilitates rapid response under pressure, thereby strengthening overall team resilience. Conversely, insufficient clarity or misalignment of roles can lead to fragmented communication, delays in decision-making, and diminished effectiveness of the TTPs deployed during critical incidents. Furthermore, their findings underline that optimal role specialization enables Blue Teams to proactively anticipate threats and implement defensive strategies tailored to their organizational strengths and capabilities. Clearly structured roles also facilitate better accountability, smoother information exchange, and more precise execution of complex cybersecurity procedures, ultimately contributing to a more robust and adaptable defensive posture.

Granåsen and Andersson [15] conducted a comprehensive cross-disciplinary analysis of team effectiveness during the Baltic Cyber Shield (BCS) Cyber Defence Exercise, a one-off predecessor of LS exercise focused on the defence of critical infrastructure against simulated cyberattacks. As a predecessor to Locked Shields, the BCS involved specialized teams—Blue, Red, White, and Green—each with clearly defined roles spanning from direct cyber defence operations to technical coordination and scoring. The study's methodology

uniquely integrated technical performance metrics, such as intrusion detection efficiency and response times, with behavioral assessments like observer reports, participant surveys, and systematic analyses of team communication and strategy execution. Their results demonstrated the critical importance of combining behavioral insights with technical metrics to achieve a more accurate and nuanced evaluation of defensive capabilities. Furthermore, their analysis highlighted how team structures, functional specialization, and intra-team dynamics significantly influenced the effectiveness of cyber defence strategies.

This combined analysis of existing research underscores the importance of systematically integrating both behavioral and technical metrics when evaluating Blue Team effectiveness. The studies by Champion et al. [14] and Granåsen and Andersson [15] illustrate that clearly defined roles and structured team dynamics critically shape the performance and adaptability of defensive TTPs. These findings validate the necessity of analyzing both team composition and behavioral factors alongside technical performance, providing a holistic approach to understanding defensive effectiveness in cybersecurity exercises. Consequently, these insights directly inform the methodology adopted in this study, reinforcing the relevance of examining team structures, role specialization, and combined assessment metrics to effectively evaluate and enhance Blue Team TTPs in live-fire cyber defence scenarios.

2.2.2 Realistic Scenario Development in Cybersecurity Exercises

Designing realistic scenarios is crucial for cyber defence exercises. Major live-fire exercises like LS explicitly emphasize realism – they simulate the full complexity of a massive cyber incident (technical attacks alongside strategic, legal, and communications challenges) using cutting-edge, up-to-date technologies [2]. These scenarios put blue teams in conditions close to real crises, e.g. defending a fictional country's critical infrastructure against coordinated attacks on power grids, finance, and media systems, as in LS 2022 [2]. Such realism ensures participants practice not only technical responses but also decision-making under pressure.

One study [16] examined what makes tabletop exercise scenarios "realistic and expedient" for industrial control system (ICS) cyber attacks. The findings suggest effective scenarios must be grounded in the current threat landscape (using tactics and threats that organizations face today) and plausible within the participants' context. In practice this means scenario storylines should mirror believable attack motives and techniques relevant to the target industry. Additionally, the scenario should be challenging yet not demoralizing – it needs to push participants out of their comfort zone but still allow them to ultimately succeed, giving a sense of empowerment and confidence. These elements help keep the exercise

engaging and ensure it yields meaningful learning outcomes.

Different formats of cybersecurity exercises approach scenario design in varying ways. In CTF competitions (often used for education or recruiting), scenarios are typically a series of standalone challenges or puzzles. By contrast, live-fire exercises (like Locked Shields or red team/blue team drills) unfold as continuous, real-time attack–defence scenarios. Research literature [17] distinguishes these formats: a CTF usually involves solving predefined vulnerabilities or problems in isolated systems, whereas a live-fire exercise simulates ongoing cyber attacks/defences on a network in real time, creating conditions much closer to an actual incident. Live-fire scenarios thus tend to offer higher fidelity – participants must respond to unfolding events and multiple attack vectors dynamically, which more closely mirrors real-world cyber operations

Holm and Sommestad [18] explored the use of automated threat emulation tools—essentially AI-driven or scripted adversaries—to create realistic, adaptive cyber attacks during exercises. Their empirical findings suggest that these automated systems can closely replicate real adversary behaviors, potentially providing training scenarios as balanced and believable as those designed by human planners. A similar approach is utilized in the Locked Shields exercise, which integrates automated attack tools alongside skilled human red teams. These human operators manually perform complex cyber attacks, complementing automated threats and adding unpredictability. By combining human expertise with automation, Locked Shields creates a challenging, high-pressure environment, enabling Blue Teams to realistically practice and enhance their defensive capabilities against sophisticated cyber threats.

2.2.3 Empirical Studies on Blue Team Effectiveness

Scoring in these high-fidelity exercises quantifies Blue Team "success" across multiple facets: for example, the LS 2022 scoring algorithm awarded points for each cyber-attack successfully defended, for maintaining service availability (uptime of critical services), and for handling forensic and legal challenges [19]. The winning team is typically the one that performs well across all categories, not just excelling in one area [15]. One analysis of U.S. regional Collegiate Cyber Defence Competitions (CCDC) in 2020 found that more experienced teams (and those with prior simulation-based training) earned significantly higher scores for keeping services running, correctly handling scenario inject tasks, and mitigating Red Team attacks [20]. Bayesian analysis showed prior team experience was a strong predictor of better service uptime and incident response performance [20].

Effective teamwork and communication are consistently identified as key enablers of strong

defensive performance. In detailed post-exercise analyses, researchers have observed that Blue Teams with clear internal communication flows and collaborative workflows respond to attacks more effectively [21]. For example, an empirical study of a multinational cyber exercise found that teams suffering communication breakdowns failed to report incidents or share updates in time, directly causing lower performance scores in incident analysis and reporting tasks [15].

Structured research on cyber defence competitions reinforces the critical role of teamwork. A 2018 analysis using the Observational Assessment of Teamwork (OAT) instrument found that effective collaboration (along with experience and role specialization) was a statistically significant predictor of a Blue Team's ability to detect and mitigate attacks in time [21]. In other words, teams that communicated well – sharing threat information, updating teammates on evolving incidents, and coordinating their defensive actions – tended to outperform less-cohesive teams. Similarly, high-performing teams often have well-defined internal roles and strong leadership facilitating communication, which together yield fast decision-making during attacks [20]. Beyond internal coordination, communication with external entities (leadership, other teams, etc.) can also matter.

2.2.4 Blue Team Evaluation in Locked Shields

Evaluating the technical skills of Blue Teams requires a combination of performance metrics and structured challenges. Live-fire cyber exercises provide a practical framework for this, as they score teams on a range of technical tasks under realistic conditions. For instance, Locked Shields uses a scoring system which covers multiple technical domains: service availability (keeping critical services up under attack), incident response and forensics (analyzing and recovering from attacks), and attack mitigation (preventing or stopping Red Team exploits). Each Blue Team's technical proficiency is continuously assessed by automated monitors – e.g., uptime bots that ping services and deduct points for downtime or misconfigurations. Figure 1 below, produced from the study by Mäses et al. [22], illustrates a typical weighting of scoring categories in Locked Shields.

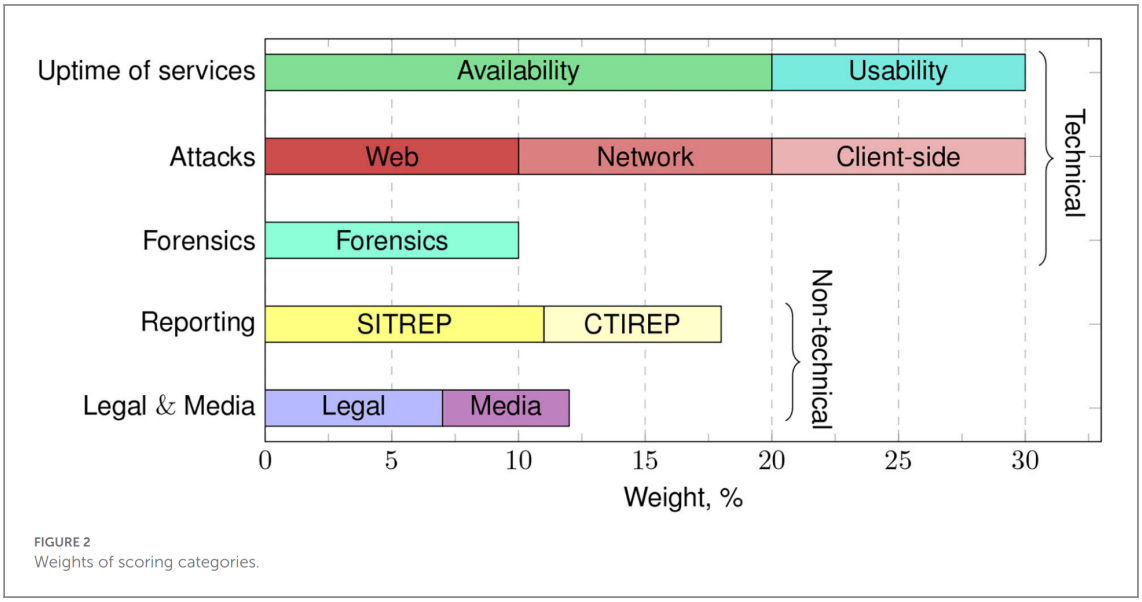

Figure 1. Example scoring distribution from a large cyber defence exercise

In Locked Shields, automated scoring bots rigorously test defensive systems and their configurations. For example, in Locked Shields the White Team's scoring engine performs regular checks on Blue Team systems: if required services or ports become unavailable, points are subtracted in real time [22]. Red Team attack success is another technical metric – every successful exploit (web, network, or client-side attack) costs the Blue Team points, creating a direct measure of how well they can defend against a certain category of threats. Blue Teams can also earn points for successful forensic investigations (solving CTF-style challenges to uncover attack evidence) and for producing high-quality incident reports [22]. The combination of these elements provides a structured assessment of a team's technical abilities: uptime and attack-defence scores reflect hardening and monitoring skills, while forensic and reporting scores reflect analytical and response capabilities.

In cybersecurity exercises like Locked Shields, scoring systems play a key role in shaping team priorities and measuring performance. The study by Mäses et al. (2020) provides insight into how scoring influences team strategies, revealing that teams often align their efforts with scoring categories, prioritizing service uptime, defending against various attack vectors, and ensuring timely reporting [22]. Recent work also explores automating the evaluation of blue-team performance in cyber-range exercises by using a report-based method to automatically assess blue-team posture, strategy and response effectiveness [23].

Granåsen and Andersson applied six different methods to measure Blue Team effectiveness in a cyber exercise, including automated service availability monitoring, network intrusion

detection log analysis, and exploratory data analysis of system events, alongside human observer reports and team surveys [15]. This approach allowed them to quantify technical performance (e.g. how many attacks were detected or defended) and also examine why certain teams performed better, by looking at team processes and structures. Their key conclusion was that no single metric suffices – only a blend of technical performance measures and behavioural assessments can holistically judge Blue Team skill levels.

2.3 Blue Team Operations

In Locked Shields, Blue Teams act as the first responders to cyber incidents. They consist of cybersecurity professionals responsible for identifying vulnerabilities within their team's infrastructure and applications [24]. By detecting and mitigating these weaknesses, Blue Teams implement security procedures and controls that strengthen system resilience and reduce the likelihood of successful attacks [24].

2.3.1 Role of Blue Teams in Locked Shields

Since Locked Shields is designed to simulate challenging cyber attack scenarios, providing blue teams with a realistic environment is essential to test their defences and response strategies. As a live-fire exercise, Locked Shields emphasizes that any vulnerabilities left unpatched can be exploited by adversaries in real time. In such cases, the exercise becomes reactive, and blue teams are forced to confront live threats as they emerge. They must quickly identify these exploited vulnerabilities and adjust their defensive measures in real time. This dynamic environment not only tests the teams' technical skills in vulnerability scanning and incident response but also their ability to adapt under pressure, mirroring the conditions of actual cyber attacks.

During the exercise, blue teams are expected to maintain the availability, integrity, and confidentiality of multiple systems and services, often simulating national critical infrastructure. In the context of Locked Shields, this includes realistic representations of power grid control systems, banking and financial platforms, military communication networks, and public service applications—all of which are essential to the functioning of a modern nation. Their technical responsibilities range from hardening operating systems and network configurations to monitoring traffic for signs of intrusion, analysing logs for indicators of compromise, and performing digital forensics on detected malware. In parallel, they must manage communication with the exercise's "management" layer—handling incident reporting, coordinating with other defensive teams, and responding to media or legal injects that test organizational and strategic decision-making.

There is also growing interest in augmenting blue team tasks with automation: for instance, in "Next Steps in Cyber Blue Team Automation—Leveraging the Power of LLMs," the authors propose using large-language models (LLMs) to automate support-ticket processing, log and network-traffic analysis, misconfiguration detection and remediation, and generation of human-readable reports [25]. Moreover, according to interview data from participants of LS 2024 — at least one Blue Team used AI to generate support-ticket responses, indicating that AI-assisted blue-team automation is already being trialed.

In summary, within the context of Locked Shields, blue teams not only focus on technical aspects such as vulnerability scanning and incident response but also play a critical role in shaping strategic defence policies. Their ability to adapt to evolving threats in a simulated environment prepares organizations for real-world cyber challenges, underscoring the significance of robust blue team operations in modern cybersecurity defence strategies.

2.3.2 Blue Team Tactics, Techniques, and Procedures (TTPs)

In cybersecurity the term TTP stands for tactics, techniques and procedures [26]. For blue teams, TTPs represent the structured set of defensive measures and operational practices that guide their efforts in securing an organization's digital environment. Drawing on the semantic framework presented in [27], blue team TTPs can be defined as follows:

- **Tactics:** These are the overarching defensive strategies and high-level goals that blue teams adopt. Tactics include initiatives such as continuous monitoring, proactive vulnerability management, rapid incident response, and the systematic hardening of systems and networks.
- **Techniques:** These refer to the specific methods and tools that blue teams utilize to implement their tactical objectives. Examples include using Security Information and Event Management (SIEM) systems for real-time log analysis, deploying Intrusion Detection/Prevention Systems (IDS/IPS), performing forensic analyses after security incidents, and conducting regular penetration testing or vulnerability scans.
- **Procedures:** These are the detailed, repeatable processes that blue teams follow to ensure consistency and efficiency in their operations. Procedures encompass activities such as the step-by-step execution of incident response, creating a new firewall rule creation request, and so on.

In the context of Locked Shields blue team TTPs become essential, because the exercise simulates real-world attack scenarios, any vulnerabilities left unpatched may be immediately exploited. As a result, blue teams must try to rapidly detect live threats, apply remediation measures, and adapt their defensive posture based on the current situation.

3. Methodology

The methodology of this research is designed to systematically evaluate the effectiveness of Blue Team TTPs used during the Locked Shields 2024 cyber defense exercise. Given the complexity of cyber operations and the multidimensional nature of defensive strategies, a mixed-methods research approach was chosen. This combination of qualitative and quantitative methods facilitates a holistic analysis, integrating in-depth qualitative insights obtained from key Blue Team members with quantitative evaluations of team performance metrics. This chapter outlines the data collection and analysis techniques, discusses the rationale behind methodological choices, and addresses the validation and limitations inherent to the research design.

3.1 Research Design

The research design of this study uses a mixed-methods approach to evaluate Blue Team TTPs utilized during Locked Shields 2024. Initially, the intention of the study was to interview and collect data from all 18 participating teams; however, a limited response rate reduced the scope of analysis to four Blue Teams, from which both quantitative and qualitative data were obtained. Quantitative data, consisting primarily of detailed exercise scores from the FER and general information submitted by teams immediately following Locked Shields 2024, provides measurable indicators of defensive effectiveness. In addition, qualitative insights were gathered through semi-structured interviews conducted approximately nine months post-exercise (between January 21 and February 28, 2025). These interviews aimed to better understand teams' experiences, preparedness, and strategic decision-making during the exercise in order to enrich and contextualize the quantitative data. By combining quantitative metrics with qualitative narratives, this research design facilitates a comprehensive evaluation of Blue Team TTP effectiveness and enables a validation of defensive strategies.

The choice of a mixed-methods design was further validated by a later study carried out by the CCDCOE in 2025, which tried to answer similar research questions using only quantitative survey methods. That study faced a very low response rate, and the data it gathered did not provide enough detail to understand the strategies or decision-making behind Blue Team actions. Without qualitative input, it was difficult to explain why certain defensive choices were made or how teams coordinated their responses during the exercise. This showed the importance of including qualitative methods, as they help capture the

human and strategic factors that cannot be seen solely through quantitative data.

3.1.1 Data Collection

This study uses a mixed-methods approach combining both quantitative and qualitative data collection to ensure a comprehensive evaluation of Blue Team effectiveness during Locked Shields 2024.

Quantitative Data Collection

The quantitative data collected from Locked Shields 2024 includes structured information provided by teams shortly after the event and detailed scoring metrics provided by the FER for each Blue Team that participated in the exercise. These metrics include:

- **Attack_CS**: Negative scoring indicating successful attacks against computer systems (higher negative scores imply poorer performance).
- **Attack_W**: Negative scoring for successful attacks against web services.
- **Attack_N**: Negative scoring reflecting successful attacks against network infrastructure.
- **STRATCOM** (Strategic Communications): Positive scoring evaluating effectiveness in strategic communication tasks.
- **USABILITY**: Positive scoring reflecting the operational usability of systems, penalizing overloading of systems with tools or tasks that degrade system performance.
- **SITREP** (Situation Reports): Positive scoring reflecting the quality and accuracy of situation reports provided by the teams.
- **REVERTS**: Negative scoring penalties applied when teams reset compromised machines to their initial state.
- **LEGAL**: Positive scoring based on the quality and timeliness of legal reports produced during the exercise.
- **FOR** (Forensics): Positive scoring reflecting performance in forensic investigation tasks.
- **Availability**: Positive scoring based on the uptime and operational availability of systems throughout the exercise.
- **CTIREP** (Cyber Threat Intelligence Reports): Positive scoring based on the quantity and assessed quality of cyber threat intelligence reports submitted during the exercise.
- **XP** The XP scoring category refers to “X-points,“ or extra points, which are discretionary points awarded or penalized by the exercise organizers.

These metrics provide objective indicators of performance, allowing for systematic quan-

titative comparisons among teams and for validating the effectiveness of specific TTPs identified through qualitative analysis.

Survey data was also gathered from the User Simulation Team (UST) members allowing them to evaluate the performance of the BTs based on their interactions with them during the exercise. They were asked four questions:

- Did you feel that your BT was effectively resolving issues?
- Was your BT friendly?
- Did you enjoy communicating and working with your BT?
- Would you recommend your BT to your co-workers?

They had to provide ratings for each of those questions on a scale of 0 (not at all) to 5 (very likely).

Qualitative Data Collection

To enrich and contextualize quantitative performance metrics, qualitative data was gathered through semi-structured interviews. Eight interviews were conducted between January 21 and February 28, 2025, approximately nine months after the Locked Shields 2024 exercise. Key personnel from five participating Blue Teams provided detailed insights regarding their experiences during the exercise with regards to the questions established in the semi-structured interview. One of the five planned interviews was not concluded due to time constraints, which resulted in incomplete data. Consequently, only the TTPs of four Blue Teams are analysed in this study.

Discussion of Interview Questions

The semi-structured interviews were designed to align closely with the research questions of this thesis. The interview questions were categorized into distinct thematic areas to facilitate systematic data collection and comprehensive analysis. A full list of the interview questions can be found in Appendix 2 - Semi-Structured Interview Questions.

General Questions Intended to establish context by capturing background information on participant roles, prior experience with Locked Shields, and overall team preparedness during the exercise.

Team Composition and Resources Examined how team structure and expertise influenced the teams' defensive strategies and outcomes.

Gathering information about TTPs Aimed to explore how Blue Teams planned, imple-

mented, and evaluated their defensive tactics, techniques, and procedures throughout the exercise.

Automation and Tools Assessed the extent to which automation influenced defensive performance, exploring both its benefits and limitations.

Adaptability Focused on the teams' capacity to adjust strategies dynamically in response to evolving threats during the exercise.

Evaluation and Lessons Learned Aimed to capture participants' reflections on their strategies' effectiveness and internal assessment procedures.

This structured approach to interview design ensures that the qualitative data collected is directly relevant to the thesis objectives, allowing for a nuanced and comprehensive understanding of Blue Team effectiveness during Locked Shields 2024.

3.2 Data Analysis

The data analysis in this study followed an approach that integrated qualitative insights with quantitative performance metrics to evaluate the effectiveness of Blue Team TTPs during Locked Shields 2024. Two analysis methods were applied: qualitative thematic analysis and comparative analysis. The qualitative analysis of semi-structured interviews was conducted using a thematic coding approach. Interview responses were systematically categorized according to a predefined coding scheme, applied where applicable (see Appendix 3 - Interview Categorization Table for the complete categorization table). This categorization scheme captured various dimensions of team performance and experience, including roles, expertise, communication effectiveness, strategic adaptability, automation usage, resource allocation, and overall preparedness. By quantifying and categorizing the relevant qualitative responses, the analysis enabled systematic comparisons with quantitative performance metrics.

3.2.1 Qualitative Analysis

Qualitative data collected from semi-structured interviews were analysed using thematic analysis. The recorded interviews were first transcribed and then systematically reviewed to identify key themes, patterns, and insights relevant to Blue Team strategies, experiences, and defensive effectiveness. This thematic analysis particularly focused on identifying:

- Patterns of team structure,

- Commonly reported effective or ineffective defensive TTPs,
- Patterns of communication, coordination, and adaptability within teams,
- Perceptions of automation tools' usefulness and limitations, and
- Patterns of team goals.

3.2.2 Comparative Analysis

A comparative analysis was performed to explore differences and commonalities across the four Blue Teams included in the study. This analysis specifically compared:

- Teams' self-reported levels of preparation and team composition versus their actual performance outcomes.
- Variations in performance related to specific TTP implementations, such as the use of automation or adaptive strategies.
- The impact of identified qualitative themes (such as adaptability and communication) on teams' quantitative scoring metrics.

By systematically comparing team experiences, strategies, and outcomes, this approach identified effective practices and potential performance-impacting factors that were either consistent across teams or specific to certain contexts.

3.2.3 Integration of Qualitative and Quantitative Results

Following individual qualitative and quantitative analyses, results were integrated using a mixed-methods approach. Integration involved cross-referencing thematic insights from qualitative analysis with results from quantitative data from the FER to validate findings and identify themes and contrasts. This mixed-method integration ensured a holistic interpretation, with qualitative insights providing explanations for quantitative patterns and quantitative results showing qualitative narratives in measurable outcomes.

3.3 Limitations of the Study

This study has several inherent limitations that should be considered when interpreting its findings. Firstly, the scope of analysis was constrained by a limited response rate, restricting the qualitative analysis to only four Blue Teams from the Locked Shields 2024 exercise. While the semi-structured interviews provided detailed insights, the limited sample size reduces the generalizability of results to all participating teams or other cybersecurity contexts. Secondly, the qualitative component relies significantly on participants' self-

reported data obtained through semi-structured interviews conducted approximately nine months after the exercise. This time gap might introduce recall bias, as participants' memories or perceptions could change or diminish over time, potentially affecting data accuracy. Additionally, practical factors such as limited resources, reliance on qualitative interview data, and time constraints associated with the scope of a master's thesis placed further limits on the depth of the analysis.

To ensure that the analysis remains meaningful and grounded despite these limitations, the following assumptions were made regarding the Blue Teams:

- Participants in the Blue Teams applied their TTPs under conditions that accurately reflect their actual operational capabilities, preparedness, and available resources.
- Data collected through interviews is considered reliable and genuinely represents the participants' actual experiences and perceptions during the Locked Shields 2024 exercise.

By explicitly defining these limitations and assumptions, the study ensures transparency and provides clear context for interpreting and applying its findings.

4. Results

This chapter presents the findings from both qualitative and quantitative data sources collected for this study. The goal is to evaluate and understand the effectiveness of Blue Team TTPs used during Locked Shields 2024. The results are structured in two main parts: qualitative insights gathered through semi-structured interviews with key Blue Team members, and quantitative metrics sourced from the official FER and structured post-exercise data.

A total of seven interviews were conducted across four Blue Teams (F, H, J, P), providing perspectives on team structures, strategic decisions, automation practices, communication, workflows, and lessons learned. These qualitative findings are supplemented by performance metrics as defined in chapter 3.1.1. These results form the empirical foundation for the evaluation and analysis chapters.

4.1 Qualitative Interview Results

This section presents a summary of the qualitative data gathered through semi-structured interviews with key players from four Blue Teams (F, H, J, P) who participated in Locked Shields 2024, providing context for the discussion chapter. A total of seven interviews were conducted: one each with members of Blue Team F and Blue Team H, three with members of Blue Blue Team J, and two with members of Blue Team P. Where multiple interviews were held with the same team, responses were cross-referenced to identify consistent themes and perspectives. These qualitative findings are categorized and key observations are pointed out in Section 4.1.1 to support comparative analysis.

Table 1 summarizes the number of interviews conducted with members from each participating Blue Team included in this study.

Table 1. Number of Interviews Conducted per Blue Team

Blue Team Identifier	**Number of Interviews**
BT_F	1
BT_H	1
BT_J	3
BT_P	2

4.1.1 Interview Results: General Questions (Q1–Q5)

This subsection provides an overview of the background and perspectives of the interview participants. Responses were collected from seven individuals representing four different Blue Teams (F, H, J, P). While most participants held leadership or strategic roles within their teams, one respondent from Blue Team P occupied a technical position during the exercise.

The participants came from a broad range of professional environments, including military and defence roles, technical IT positions, cybersecurity-specific roles, and managerial positions. Most of them had previously participated in Locked Shields, with the vast majority having attended the exercise at least twice. This suggests a level of familiarity with both the format and expected challenges of the event, providing a degree of experiential credibility to their insights.

Regarding their team's overall experience during Locked Shields 2024, the majority of respondents expressed a positive sentiment. Five participants described the experience in favourable terms, while two were neutral. No respondents characterized their experience negatively, indicating a generally successful and well-received engagement across the interviewed teams.

The final general question (Q5) asked participants to identify the most significant challenges their teams faced during the exercise. Unlike the background-oriented questions above, this can be used for deeper analysis in the following sections, as it touches directly on operational and organizational pain points. Figure 2 categorizes the reported challenges into five thematic groups: communication issues, time management constraints, training or skill gaps, coordination and logistics problems, and technical difficulties.

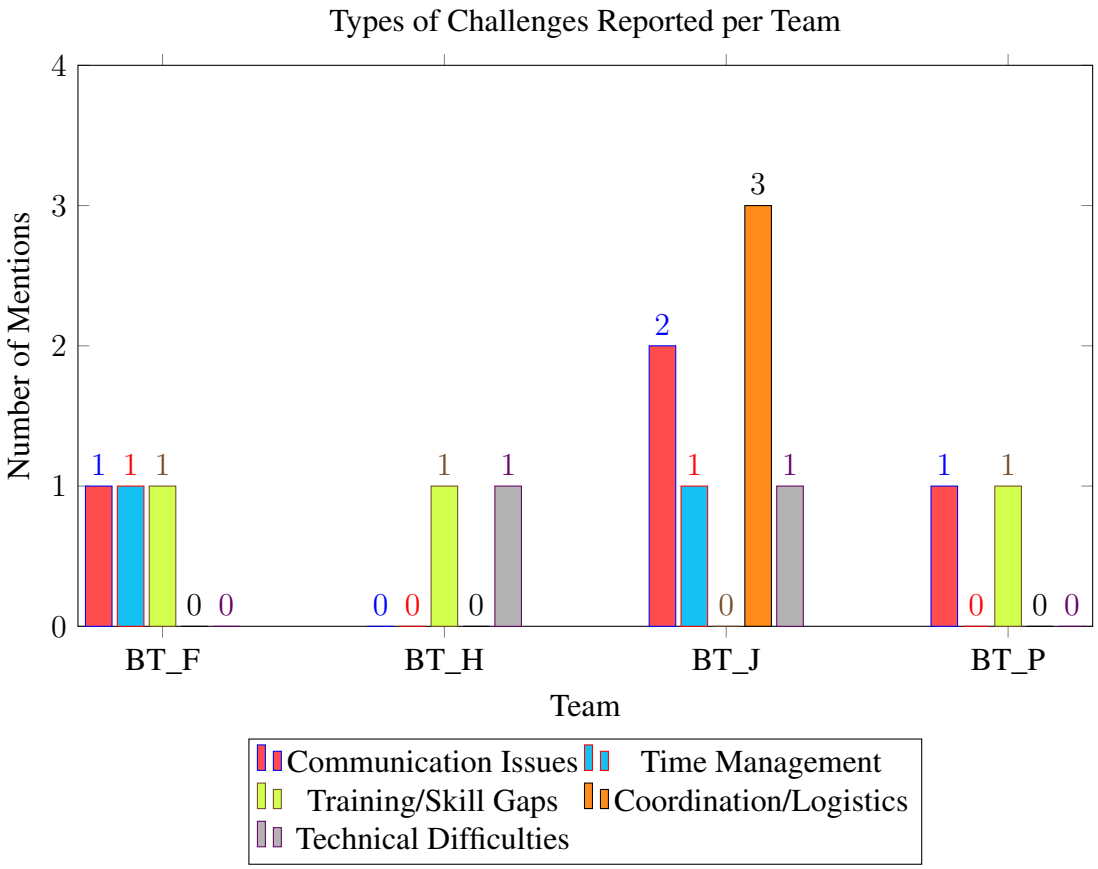

Figure 2. Categories of challenges reported by each Blue Team (aggregated across interviews).

Observations:

- **O1:** **BT_F** Reported *Communication Issues*, *Time Management*, and *Training/Skill Gaps*, suggesting organizational and skill-related concerns.
- **O2:** **BT_H** Reported *Training/Skill Gaps* and *Technical Difficulties* but did not mention communication or coordination problems, in contrast to other teams.
- **O3:** **BT_J** Reported the widest variety of issues, with all interviews referencing *Coordination/Logistics* and some mentioning *Communication Issues* or *Technical Difficulties*. In contrast, the other teams did not highlight coordination as a primary concern.
- **O4:** **BT_P** Mentioned *Communication Issues* and *Training/Skill Gaps*, which aligns with the challenges faced also by BT_F, however they did not mention problems related to time management.

4.1.2 Interview Results: Team Composition and Resources (Q6–Q9)

This section presents findings on how teams were composed and resourced for Locked Shields 2024. It covers structural organization, technical expertise across key domains, the proportion of new versus veteran members, and how team size influenced operational effectiveness.

Q6. Team Structure Type

(Original Question: "How was your team structured?")

In the context of the Locked Shields exercise, two primary team structuring approaches were identified: zone-based structures and functional structures.

A zone-based team structure organizes defenders according to the network topology created for the exercise. Typically, the network is divided into separate zones, representing distinct operational domains within the exercise, such as demilitarized zone (DMZ), 4G/5G network, Simulated Internet (SINET) or local domains. A given zone may have several different types of assets to defend, for example the DMZ zone in Locked Shields 2024 had a Windows Server, multiple flavours of Linux servers, Kubernetes and a firewall appliance. Therefore, in the zone-based team structure, sub teams are comprised of team members with different technical backgrounds and are then assigned responsibility for the entirety of a given zone. The key focus is on holistic protection of a specific area of the network.

In contrast, a functional team structure organizes defenders based on technical specialization. Sub teams are formed according to the type of asset or technology, such as Windows, Linux, web, or network infrastructure. Functional teams are responsible for their respective technologies across all zones, meaning that defenders must coordinate efforts across different domains but focus on a particular type of system or service.

Table 2 below depicts the reported team structures for each blue team.

Table 2. Reported Team Structures for Each Blue Team

Team	Team Structure
BT_F	Zone-based
BT_H	Functional
BT_J	Functional
BT_P	Functional

Observations:

O5: One team (**BT_F**) reported a *Zone-based* structure, while the other three (**BT_H**, **BT_P** and **BT_J**) used a *Functional* model.

O6: Overall, the results indicate a difference in organizational styles, showing that they took varied approaches to dividing tasks, delegating authority, and coordinating their defensive efforts during the exercise.

Q7. Expertise Composition

(Original Question: "What was the level of expertise in administration, forensics, incident response, automization, threat analysis, and pentesting within your team?")

Participants provided their responses in a qualitative manner, describing the general level of expertise within their teams rather than assigning explicit numerical values. To enable comparative analysis, these qualitative assessments were subsequently translated into a quantitative range from 0 to 5, based on the expressed level of confidence and proficiency conveyed in each response. Missing or unclear data points were represented as 0. The self-reported expertise ratings across the core domains are visualized in Figure 3, with missing or unclear data points represented as 0.

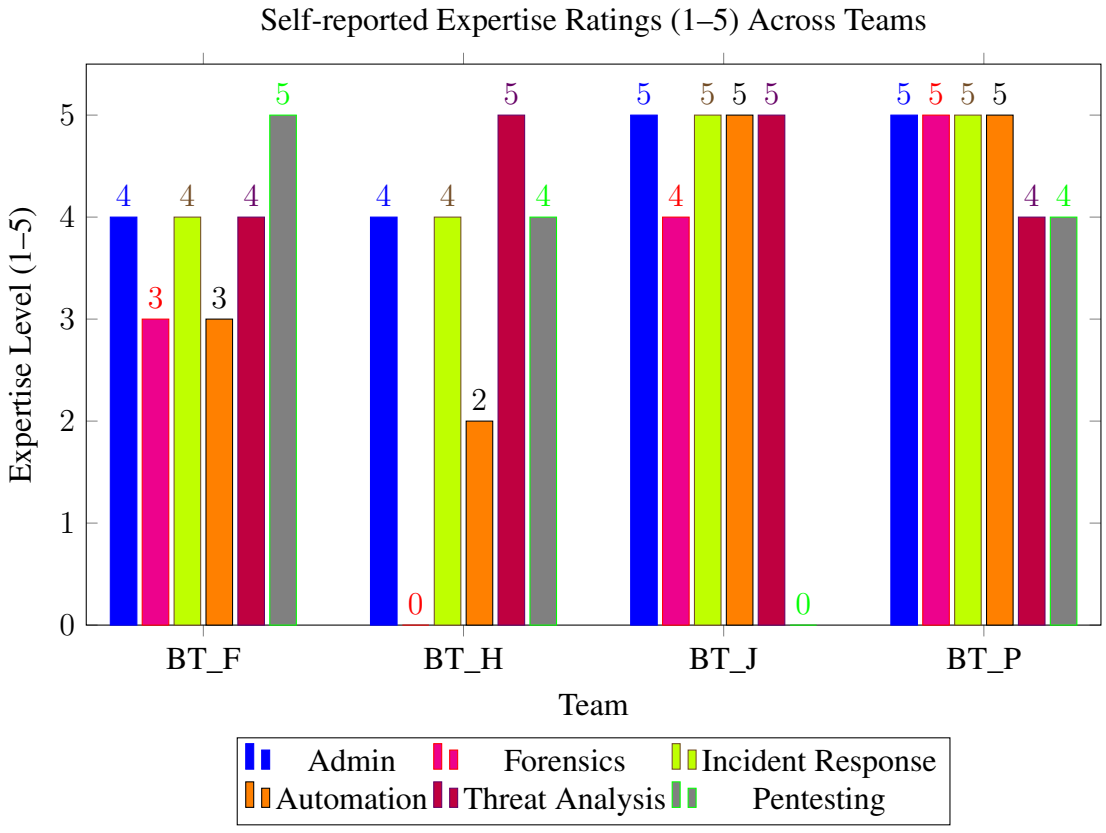

Figure 3. Expertise levels self-reported by each Blue Team across key domains.

Observations:

- **O7:** **BT_F** reported moderate to high ratings (3–5) in all domains, peaking at 5 in *Pentesting*.
- **O8:** **BT_H** demonstrated advanced *Threat Analysis* (5), while *Automation* (2) and *Forensics* (0 = unclear) were lower or missing respectively.
- **O9:** **BT_J** (three interviews) showed unanimous *5* across *Admin, Incident Response, Automation, Threat Analysis*, with *Forensics* at 4. All interviewees left *Pentesting* unclear (0).
- **O10:** **BT_P** (two interviews) had one participant providing 4–5 in each area and another who left all ratings unclear, due to not having direct communication with the other subteams.
- **O11:** Across teams, *Threat Analysis* appears consistently high, whereas *Pentesting* and *Forensics* have data gaps. This suggests strong detection capabilities but partial or uncertain offensive skill sets.

Q7.1 Team Composition by Overall Expertise Level

Table 3 presents an overview of the overall expertise levels within the teams interviewed, as assessed through participant responses.

Table 3. Overall Team Expertise Composition

Interview	Team	Overall Expertise
BT_F_1	BT_F	Medium
BT_H_1	BT_H	High
BT_J_1	BT_J	High
BT_J_2	BT_J	High
BT_J_3	BT_J	High
BT_P_1	BT_P	High
BT_P_2	BT_P	Unclear

Observations:

- **O12:** The majority of interviews (5 out of 7) described their teams as having a *High* level of expertise.
- **O13:** Only **BT_F_1** provided a *Medium* assessment, suggesting a more balanced view of individual skill sets.
- **O14:** One respondent (**BT_P_2**) marked *Unclear*, reflecting uncertainty about overall team composition.

O15: Overall, these results indicate that most participants perceive their teams as highly skilled.

Q8. New vs. Veteran Composition

Original Question: "What was the proportion of new vs veteran individuals in the team?" Table 4 presents the distribution of new and veteran team members as reported during the interview.

Table 4. Self-Reported Percentage of New Members in Each Blue Team

Interview	Team	% New (Approx.)
BT_F_1	BT_F	30–50
BT_H_1	BT_H	60–70
BT_J_1	BT_J	30–40
BT_J_2	BT_J	30–40
BT_J_3	BT_J	30–40
BT_P_1	BT_P	30
BT_P_2	BT_P	50

Observations:

O16: **BT_F** had an estimated 30–50% new members, indicating a balanced mix of newcomers and veterans.

O17: **BT_H** reported the highest proportion of new members (60–70%), suggesting a comparatively smaller amount of veteran members.

O18: **BT_J** showed consistent estimates (30–40%) across three interviews, pointing to internal agreement on the team's composition.

O19: **BT_P** (two interviews) provided different estimates of new members, with one interviewee citing about one third (33%) and the other mentioning 50%. The latter also reported uncertainty regarding other aspects of team composition, suggesting that the actual figure may vary.

O20: Overall, while most teams had between 30% and 50% newcomers, **BT_H** stood out with a significantly higher figure, potentially affecting the team's experience level.

Q9. Reported Team-Size Adequacy

Original question: "Did the size of your team impact your ability to execute defensive strategies effectively?"

Interviewees' assessments of team size were grouped into three categories: *Understaffed*, *Optimal*, and *Overstaffed*. Figure 4 summarises these perceptions for each Blue Team.

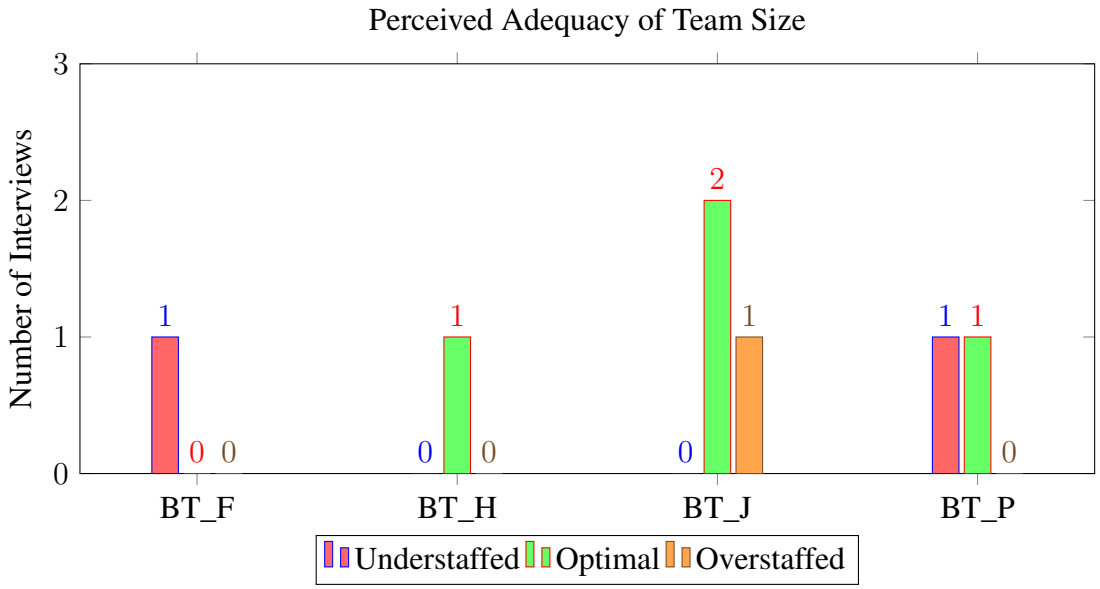

Figure 4. Interviewee sentiment on whether their team size was *Understaffed*, *Optimal*, or *Overstaffed*.

Observations

O21: **BT_F** assessed their team as *understaffed*. The interviewee noted that the team could have benefited from as many as thirty additional members.

O22: **BT_H** reported their team size as *optimal*. The interviewee emphasized that a smaller, more focused team is beneficial rather than simply increasing numbers.

O23: **BT_J** had mixed responses. Two interviewees judged the team size to be *optimal*. A third interviewee described the team as *overstaffed*, citing challenges such as limited physical space and communication difficulties.

O24: **BT_P** also presented mixed views. One interviewee considered the team size *optimal*. Another, occupying a technical role, characterized the team as *understaffed*. The latter explained that not every host system had a dedicated defender, and team members had to rotate between machines.

4.1.3 Interview Results: Gathering Information About TTPs (Q10–Q16)

Q10. Planned TTPs

(Original Question: "Can you describe your planned TTPs?")
The planned TTPs have been grouped into three categories—*Preventive*, *Detective*, *Responsive* based on their primary function or intent. This categorization provides a structured view of each team's approach.

- **BT_F**
  - *Preventive*:
    - Web application sandboxing
    - Removal of external dependencies ("de-CDNing")
  - *Detective*:
    - Deployment of EDR agents on endpoints
    - Use of endpoint monitoring agents
  - *Responsive*:
    - Dedicated reporting and SOC subteams
    - Reporting subteam managed the ticketing process
    - Incident response conducted ad-hoc based on role and availability
- **BT_H**
  - *Preventive*:
    - Proactive reconnaissance during familiarization days; vulnerabilities were documented
    - Predefined Day 1 action plan
  - *Detective*:
    - Endpoint monitoring agents
    - Network traffic analysis
    - Deployment of EDR agents
  - *Responsive*:
    - Use of threat hunting teams to identify root causes and persistent threats
    - Incident coordinators managed ticketing and communication with reporting teams
    - Internal ticketing system
- **BT_J**
  - *Preventive*:
    - Tactical plans for each subteam
    - Centralized code and tool repository

        - Introductory guide for team members
        - Detailed plan for the first 30 + 30 minutes of the exercise
        - Partial operating system reinstalls
        - Hardening of operating systems and firewalls
        - Proactive reconnaissance during familiarization days
        - Scraping and dockerization of web applications
        - Deployment of web application firewall
    - *Detective*:
        - Endpoint monitoring agents
        - Deployment of EDR agents
    - *Responsive*:
        - Ticketing system integrated with custom chatbots for forwarding threat-hunting data
        - Coordinated incident response procedures
        - Internal ticketing system
        - Shared network map documenting roles and responsibilities
- **BT_P**
    - *Preventive*:
        - Hardening of operating systems and firewalls
        - Specialized support team developed integrations and automations
        - Introductory guide for team members
        - Detailed plan for the first 30 minutes
        - Comprehensive review of exercise documentation; responsibilities were distributed by rule
    - *Detective*:
        - Endpoint monitoring agents
        - Deployment of EDR agents
    - *Responsive*:
        - Wiki-based documentation
        - Defined reporting procedures
        - Defined escalation procedures

Observations:

O25: **BT_F** used web application sandboxing and removal of external dependencies as preventive controls. Detection relied on EDR and endpoint monitoring agents. Response involved dedicated SOC and reporting subteams, though incident handling was largely ad hoc.

O26: **BT_H** focused on early reconnaissance and a predefined Day 1 plan. Detection

combined endpoint and network monitoring with EDR tools. Response was structured by using an internal ticketing system, also threat-hunting teams and incident coordinators were utilized for reporting and triage.

O27: **BT_J** demonstrated comprehensive planning, including subteam tactical plans, host hardening, and containerization of web applications. Detection included EDR and endpoint agents. Response was highly coordinated via internal ticketing systems, chatbots, and shared network documentation.

O28: **BT_P** prioritized host hardening and distributed responsibilities based on detailed documentation analysis. Detection involved standard EDR and monitoring agents. Response processes were documented in a team wiki, with clearly defined reporting and escalation procedures.

O29: **Cross-team Similarities**: All teams implemented host hardening, EDR and endpoint monitoring, and had some form of response documentation or coordination.

O30: **Cross-team Differences**: Team J showed the highest degree of integration and automation. Team F relied on flexible, ad hoc response. Team H emphasized early planning and role-based coordination, while Team P focused on documentation and procedural clarity.

Q11. Measurable Impact of Defensive Actions

(Original Question: "Were there specific defensive actions that had a measurable impact on system uptime or other performance metrics? Why do you think so, how did you measure?")

BT_F Sandboxing web applications led to visibly improved uptime scores, as attackers were unable to bring down any webservers. Although the team spent less effort patching the actual applications themselves—leaving them exposed to defacement, XSS and data-leak attacks.

BT_H Deployed GPOs caused availability issues such as RDP being blocked on certain endpoints. The team also cross-checked threats on other machines to improve reporting scores.

BT_J Due to a communication error an automation script meant to prevent user-side attacks was never run, allowing Phase 1 user-side attacks to go through, which made them score penalty points. Firewalling outbound traffic on a large amount of endpoints prevented malware beacons from reaching the attacker C2 servers. A RT attack deleted a key DLL file on Windows endpoint leading to usability score reductions.

BT_P Web Application Firewall (WAF) over-filtered connections, leading to usability score reductions.

Observations:

- **O31:** **BT_F** observed better uptime from web application sandboxing, as attackers could not bring web servers down.
- **O32:** **BT_F** did not patch the source code of the web applications leading to defacement, XSS and data-leak attacks.
- **O33:** **BT_H** improved reporting scores by checking whether similar attacks were occurring on other machines.
- **O34:** **BT_J** faced a penalty in score in computer system attack category when a crucial automation script was never run, allowing user-side attacks in Phase 1.
- **O35:** **BT_P** Web Application Firewall (WAF) over-filtered connections, leading to usability score reductions.

Q12. Resource Allocation for Multiple Attack Vectors

(Original Question: "How did you allocate resources such as time and team members to address multiple attack vectors?")

BT_F Team members were assigned to specific groups of machines that they were responsible for, however when UST tickets started coming in then they would allocate team members ad-hoc to provide information for reporting.

BT_H Allocation of time and team members was decentralized, with sub team leads managing tasks independently. Team members were encouraged to be proactive and solve the problem as efficiently as possible. This was done by allocating 5-15 minutes per problem, if the issue could not be resolved by that time then it should be escalated.

BT_J Time and personnel allocation was primarily managed by subteam leads. A specialized "ambulance" subgroup, composed of highly skilled members, was designated to assist with escalated or complex tasks. Regular rehearsals conducted prior to the exercise contributed to effective time management during operations.

BT_P Team members were initially assigned to specific hosts and tasks; however, those who completed their assignments early were reallocated by team leads to support more active areas. A dedicated subteam was responsible for overseeing the management of ongoing tasks and incidents.

Observations:

- **O36:** **BT_F** and **BT_P** assigned members to specific machine groups and reassigned people on an ad-hoc basis
- **O37:** **BT_H** adopted a decentralized scheme, with sub team leads overseeing tasks and imposing a short time limit (5–15 minutes) before escalating issues to senior personnel.
- **O38:** **BT_J** structured resource allocation around sub team leads and an "ambulance" subgroup of highly skilled responders
- **O39:** **BT_P** dedicated a sub team that was responsible for overseeing the active tasks and incidents.

Q13. Communication and Coordination Under Pressure

(Original Question: "How did your team communicate and coordinate tasks during high-pressure moments?")

BT_F Developed a custom bot in a chat server that could open tickets and channels for reporting team. The reporting team would handle the tickets and query team members for technical information when necessary.

BT_H Adopted a 5-15 minute rule for problem-solving, escalating to more senior members if issues could not be solved in that time frame. A digital communication system was used to communicate issues within the team.

BT_J Hybrid communications was used, with some part of the team being on-site and the rest of the team was using a digital communication channel. If a sub team could not internally solve a problem, it was elevated to ambulance team. Custom chat bot based ticketing system was developed to handle incidents internally. Dedicated help desk was created for handling UST tickets.

BT_P The entire team operated fully on-site, complemented by chat servers. Dedicated personnel was dealing with support tickets.

Observations:

- **O40:** **BT_F** relied on a chat-based workflow with an automated bot to open tickets.
- **O41:** **BT_F** set up dedicated channels for reporting sub team, which handled both internal and UST tickets

O42: **BT_H** used a 5–15 minute rule for problem-solving. If a sub team could not resolve an issue within that window, it was escalated

O43: **BT_J** adopted a hybrid model, with some members on-site and others using digital communication.

O44: **BT_J** issues were escalated to an "ambulance team" of experts

O45: **BT_P** was fully on-site but also used chat servers for information sharing

O46: All teams mentioned having dedicated personnel for handling UST tickets.

Q14. Ineffective TTPs

(Original Question: "Were there any TTPs your team had planned that turned out to be ineffective?")

BT_F They used a multi purpose proxy to provide a WAF filter and dynamically rewrite CDN connections. The WAF did not provide the expected level of protection; no concrete examples of successful filtering were noted. They shifted toward using the proxy just for rewriting CDN connections.

BT_H Launch plan did not go as expected, as some of tooling was not economical enough and therefore degraded performance in the environment. It was also difficult to correlate information that was coming in from all the different tooling that was deployed.

BT_J Communication pipelines between the technical and non-technical parts could be improved.

BT_P References to remote resources were not replaced in web application source code even though it was previously agreed-upon.

Observations:

O47: **BT_F**'s multi-purpose proxy's WAF functionality did not meet expectations, prompting a shift to using the proxy primarily for rewriting CDN connections.

O48: **BT_H** experienced challenges during launch due to inefficient tooling, which negatively impacted system performance

O49: **BT_J** noted that communication between technical and non-technical components needed improvement. The existing pipelines did not fully bridge the gap.

O50: **BT_P** Certain preplanned actions, such as removing references to remote resources in the web application code, were never completed, suggesting a shortfall in following through on preparations.

Q15. Pre-Exercise Testing and Preparation

(Original Question: "Were you able to test and prepare the TTPs prior to the exercise?")

BT_F Hosted typical PHP applications in custom test environments and focused on developing core automations for the initial phase of the exercise. They used knowledge from previous exercises but intentionally avoided reusing entire old solutions to maintain in their words "the spirit of the exercise".

BT_H Focused on familiarizing new members to scripts and tooling in custom test environments. Created a centralized knowledge repository.

BT_J Conducted multiple "hackathons" and hosted weekly sub team meetings prior to the exercise. Multiple full-team rehearsals were had day 0. Developed custom test environments for testing out automation and tooling.

BT_P Ran two large internal exercises for rehearsing, one purely technical and one full-team simulation covering reporting and other non-technical areas.

Observations:

- **O51:** **BT_F** tested common PHP applications in custom test environments and developed automations.
- **O52:** **BT_F** avoided reusing entire solutions from previous year exercises to preserve "the spirit of the exercise."
- **O53:** **BT_H** practised scripting and tooling in custom test environments and created a centralized knowledge repository.
- **O54:** **BT_J** hosted multiple hackathons and weekly sub team meetings.
- **O55:** **BT_J** also held multiple Day 0 rehearsals in order to test their tooling.
- **O56:** **BT_P** organized two large internal exercises—one focused on technical defence and automation, and another covering full-team processes including reporting and coordination.

Q16. Team Preparedness

(Original Question: "How would you rate your team's preparedness for the exercise?")

Interviewee responses were rated on a scale from 0 to 5, where 5 indicated a very high level of preparedness and 0 indicated no preparedness at all. Figure 5 shows the team preparedness rating.

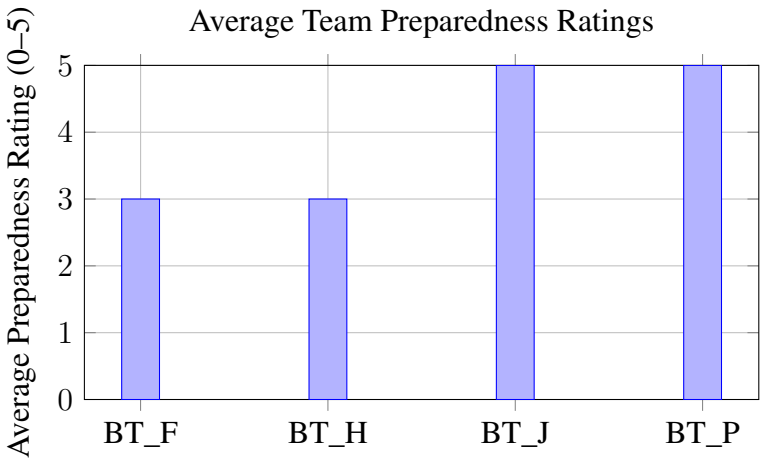

Figure 5. Average preparedness ratings (0–5) reported by Blue Teams. All team members from the same team gave identical ratings.

Observations:

O57: **BT_F** noted a lack of time among team members to prepare prior to the exercise. Overall, they felt they prepared on an "okay" level, while improvements were possible had they started earlier.

O58: **BT_H** described their readiness as "okay" overall. It was explained that the team had proactive volunteers who did not fully grasp the scale and depth of Locked Shields.

O59: **BT_J** claimed a "very good" level of preparedness thanks to motivated volunteers and detailed planning on the sub team level. .

O60: **BT_P** claimed they were "very prepared", due to the extensive preparations and training

4.1.4 Interview Results: Automation and Tools (Q17–Q19)

This subsection presents the tools and automation strategies used by the Blue Teams, based on responses to Questions 17 through 19.

Q17. Automated Tools and Processes

(Original Question: "Describe the automated tools or processes that your team used.")

Table 5 summarizes the automation and tooling methods reported by each Blue Team during the exercise.

Table 5. Automation and Tooling Methods Reported by Blue Teams

Team	Automation Tools	Scripting	Custom Tooling	Discord Bots	ELK Stack Automations
BT_F	✓	✓		✓	
BT_H	✓	✓			
BT_J	✓	✓	✓	✓	✓
BT_P	✓	✓	✓		

O61: All Blue Teams reported using both scripting and automation tools, indicating a baseline reliance on automated task execution and custom script development.

O62: Custom tooling was used by BT_J and BT_P, suggesting a more advanced or tailored automation environment in these teams.

O63: BT_J was the only team to report using all listed automation and tooling methods, including Discord bots and ELK Stack automations, reflecting the most comprehensive integration of automation.

O64: BT_F and BT_J reported the use of Discord bots, highlighting their use of communication-integrated automation to support coordination or reporting tasks.

O65: ELK Stack automations were used exclusively by BT_J, suggesting a higher level of integration between security monitoring and automated workflows in this team.

Q18. Balancing Automation and Manual Tasks

(Original Question: "How much time did you spend on automation vs manual tasks?")

Observations: BT_F

O66: The core deployment was automated, while subsequent tasks were performed manually. This indicates that the team used automation primarily for initial system provisioning and configuration management.

BT_H

O67: Automation was centralized within the team, indicating that dedicated members or subgroups were responsible for developing, maintaining, and executing automated tasks.

O68: The teams approach was to balance automated and manual processes, using automation for repetitive tasks while reserving manual work for analysis and decision-making.

BT_J

O69: Most of the preparation for the exercise went into automation tools.
O70: All of the web services were automatically scraped and put into docker containers.
O71: Specialized systems engineers used a manual approach, as automating those tasks was considered too complex.

BT_P

O72: Tried to push automation as much as feasible.
O73: During execution the first 30 minutes was automated the rest was manual.

Q19. Automation Limitations and Manual Intervention

(Original Question: "Were there scenarios where automation failed or where manual intervention was necessary?")

Observations:

O74: **BT_F** Automation generally worked, some cases required manual intervention

BT_H

O75: Certain tasks, such as identifying persistence mechanisms, could not be automated and required manual inspection of the web servers.

BT_J

O76: Ansible playbooks were used from previous years and since the environment had changed some scripts failed to run. These shortcomings were fixed during the rehearsal.
O77: Connectivity issues during automation process caused it to not finish.
O78: Automated installation of monitoring agents did not work as expected due to performance constraints on the VMs.

BT_P

O79: Planned for manual intervention in scenarios where automation was expected to be unreliable or insufficient.

O80: Encountered cases where system changes between Day 0 and Day 1 led to automation failures.

O81: Experienced connectivity issues during the automation process, which were related to network hardening activities.

4.1.5 Interview Results: Adaptability (Q20–Q21)

This subsection summarizes how Blue Teams adapted to unexpected challenges and shifting conditions during the exercise, based on responses to Questions 20 and 21.

Q20. Abandoning or Shifting TTPs

(Original Question: "Were there moments where your team had to abandon one TTP in favor of another? What prompted this shift?")

BT_F

O82: The team observed that the WAF failed to provide meaningful protection, leading them to shift toward alternative defensive measures.

BT_H

O83: Shifted priorities during the exercise by reassigning people when problems occurred.

BT_J

O84: The network team had to modify the firewall rule request procedure after it became overcrowded and inefficient.

O85: TTPs were shifted towards optimizing uptime of machines and services.

BT_P

O86: Nothing significant had to be changed, everything worked well and only minor issues appeared.

Q21. Role of Improvisation

(Original Question: "Can you share an example where improvisation played a key role in your defensive actions?")

BT_F

O87: Leadership decided to allocate more members to the forensics sub team during the execution.

O88: Incident response was improvised to some degree – they had general ideas, but had to go off of availability of team members.

BT_H

O89: One of the firewalls got compromised by the opposing Red Team, but the internal Red Team supporting the Blue Team managed to regain access and control of the firewall.

BT_J

O90: They managed to gain control of one of the Red Team's machines.

O91: When the Red Team took control of a Linux endpoint, a team member still had an active shell session open in the background, which allowed them to re-establish control over the compromised machine.

O92: Constantly improvised and adapted to the Red Team's actions throughout the exercise.

O93: Creating a custom "washing machine" firewall solution for special systems.

BT_P

O94: The team member responsible for UST communication used a large language model (LLM) to quickly generate responses.

O95: Lost one host during the exercise, but had a backup of the application, which was successfully restored on another machine. The team then updated the DNS entries to redirect traffic to the new host.

4.1.6 Interview Results: Evaluation and Lessons Learned (Q22–Q26)

This section reflects on the Blue Teams' self-evaluations and the key lessons they identified following the exercise, based on responses to Questions 27 through 31.

Q22. Internal Goals and Achievement Strategies

(Original Question: "Did you have any internal goals, how did you plan to achieve them?")

BT_F

O96: The general goal was to educate and facilitate knowledge transfer from experienced members to newer participants.
O97: The side goal was to win the exercise.

BT_H

O98: To bring people together from different agencies and organisations and give them the experience of Locked Shields
O99: Providing the Locked Shields experience to as many people as possible.
O100: Prioritize learning within the team.

BT_J

O101: The general goal was to win and to do as best as you can.
O102: Aimed to enhance cooperation between the military and civilian sectors, which was reflected in the composition of the mixed team.
O103: Train military and leadership roles.
O104: Enhance cooperation with allied nations.

BT_P

O105: General goal was to win and create an approach for best performance.
O106: One of the internal goals was to prepare for the next year's exercise.

Q23. Goal Effectiveness and Rationale

(Original Question: "Do you think you were effective in achieving those goals? Why do you think so?")

BT_F

O107: Yes, the team was effective in achieving its goals, as participants gained valuable experience from taking part in the Locked Shields exercise.

BT_H

O108: Yes, definitely. A lot of people got to experience the chaotic environment, got the experience about how to act in a large scale crisis.

BT_J

O109: Yes, definitely.
O110: Although the team did not achieve first place, it successfully met its other internal goals and performance objectives.
O111: Team members said they were thankful for being part of the experience.

BT_P

O112: Yes, the team achieved first place and gained numerous valuable lessons from the experience.

Q24. Significant Takeaways for Future Scenarios

(Original Question: "What were the most significant takeaways from the exercise that you would apply in future scenarios?")

BT_F

O113: Avoid aiming for a single big solution that solves 100% of the problem, instead develop a toolbox that solves 80-90% of the problem.

BT_H

O114: Gained experience with various tooling.
O115: Emphasized the importance of sharing knowledge and expertise among team members.

BT_J

O116: Recognized the need to begin the funding process earlier for future exercises.

O117: Confirmed that thorough preparation and automation greatly enhance overall performance.

O118: Identified continuous training and pre-exercise testing of plans as key factors for improving readiness and efficiency.

BT_P

O119: In exercises like Locked Shields, focusing only on cybersecurity experts—like threat hunters or analysts—can miss the bigger picture. If you don't have a solid IT operations team to actually secure and maintain the systems, you just end up with a long list of threats and no time to fix them. Good operations come first; cyber should build on top of that.

O120: Clearly defined communication pipelines are very important during large scale exercises.

Q25. Internal Lessons Learned Process

(Original Question: "Did you have an internal lessons learned process?")

BT_F

- Unclear

BT_H

- Yes, an after-action form was distributed to collect general feedback, and some team members also provided verbal input.

BT_J

- Yes, an internal reflection meeting was held after the exercise ended.
- Team members discussed their personal goals and reflected on whether they had achieved them.
- A team-level feedback process was conducted to evaluate overall performance and identify areas for improvement.

BT_P

- Yes, a follow-up meeting was held several days or weeks after the exercise, during which team leads presented their lessons learned.
- Each team member documented observations made during the exercise and reflected on key lessons learned from the experience.

Observations:

O121: BT_H: Distributed an after-action form to gather general feedback

O122: BT_J: Conducted a post-exercise reflection meeting, focusing on both individual and team goals.

O123: BT_P: Held a dedicated review session some days or weeks after the event, during which team leads presented their insights.

Q26. Areas for Improvement

(Original Question: "In hindsight, were there areas where additional preparation or resources could have improved the performance of your team?")

BT_F

- Additional preparation work could have improved overall performance.
- Some team members lacked sufficient background knowledge about the exercise.

BT_H

- The team could have benefited from having more skilled personnel on the web services side.
- Team leaders found it challenging to balance technical responsibilities with managerial duties.

BT_J

- Desired more training opportunities prior to the exercise, but these were not feasible within the available budget.
- Identified the need to include team members with expertise in artificial intelligence and machine learning.
- Recommended scheduling planning conferences earlier to allow more thorough

BT_P

- Having more participants would have allowed each host to be covered by a dedicated team member, eliminating the need to switch between systems.
- Additional preparation could have improved overall readiness and efficiency.
- A dedicated support team for handling UST tickets would have streamlined communication and reduced response delays.

Observations:

O124: **BT_F** identified insufficient preparation and limited familiarity with the exercise context as key areas for improvement.

O125: **BT_H** noted a skills gap on the web security side and highlighted difficulties in managing both technical and leadership responsibilities concurrently.

O126: **BT_J** expressed a desire for additional pre-exercise training and earlier planning opportunities, and suggested including AI/ML expertise in future team composition.

O127: **BT_P** emphasized the need for greater team size to ensure one-to-one host coverage, additional preparation time, and a dedicated support team for handling UST tickets.

4.2 Quantitative Results

This subsection outlines the quantitative performance results of the Blue Teams that participated in the Locked Shields 2024, based on data from the official final exercise report. The purpose of this section is to provide a clear overview of each team's performance. The performance of each team was scored in the following categories: ATTACK_CS, ATTACK_N, ATTACK_W, AVAILABILITY, CTIREP, FORENSICS, LEGAL, REVERTS, SITREP, STRATCOM, USABILITY and XP. The TOTAL score is formed by the sum of all categories. AVG depicts the average score across all of the 18 blue teams that participated in Locked Shields 2024. Table 6 below shows the quantitative performance data for each team.

Table 6. Blue Team Performance Scores in Locked Shields 2024

TEAM	ATTACK _CS	ATTACK _N	ATTACK _W	AVAIL ABILITY	CTI REP	FORE NSICS	LEGAL	REV ERTS	SITREP	STRAT COM	TOTAL	USABI LITY	XP
F	-2917	-1294	-2150	16368	5868	6210	9511	-350	8846	4789	54003	9120	0
H	-4375	-1868	-3367	14908	6154	8406	14505	0	7628	4667	55069	8411	0
J	-1887	-1725	-679	15634	5970	7989	12158	0	10297	4851	62527	9752	167
P	-2402	-1294	-962	16463	4792	8518	14478	-350	9197	4667	63053	10371	-425
AVG	-3703	-2659	-2576	15420	5347	7608	12226	-350	8904	5192	54246	8667	-44

Observations:

O128: **BT_H**, **BT_F** and **BT_J** all achieved high scores in CTIREP compared to **BT_-P** , indicating strong performance in communication and documentation under pressure.

O129: **BT_J** and **BT_P** achieved high total scores among all Blue Teams with a difference of only 500 points, reflecting effective and balanced performance across evaluated categories.

O130: **BT_P** received the highest usability score, indicating that the team was able to maintain service functionality and end-user experience even after hardening.

O131: **BT_H** got high penalties in the `ATTACK_CS` and `ATTACK_W` categories, possibly reflecting weaknesses in defending against specific red team tactics.

O132: **BT_P** and **BT_F** scored high in service availability, indicating effective hardening and recovery strategies.

O133: **BT_H** and **BT_P** received high scores in `LEGAL` category in contrast to the other teams.

O134: **BT_J** and **BT_P** received low penalty scores in `ATTACK_W` category in contrast to the other teams.

O135: **BT_J** and **BT_P** received high scores in `USABILITY` category in contrast to the other teams.

O136: The total score comparison shows that **BT_J** and **BT_P** significantly outperformed **BT_F** and **BT_H**. This suggests that while all teams demonstrated individual strengths, only some succeeded in achieving a consistent, high-level performance across all evaluated categories.

Promptly after the Locked Shields 2024 exercise the User Simulation Team also provided qualitative feedback on their interactions with each Blue Team. This score reflects the perceived professionalism, responsiveness, and technical competence of the Blue Teams from the perspective of UST during the exercise. Table 7 shows the feedback that was reported from the UST.

Table 7. UST Experience Score

Team	Score
F	17.13
H	12.43
J	17.00
P	13.88

Observations:

O137: **BT_F** and **BT_J** received significantly better feedback from UST team than **BT_H**

and **BT_P**.

Figure 6 shows the team sizes that were reported by the Blue Teams as input to the FER.

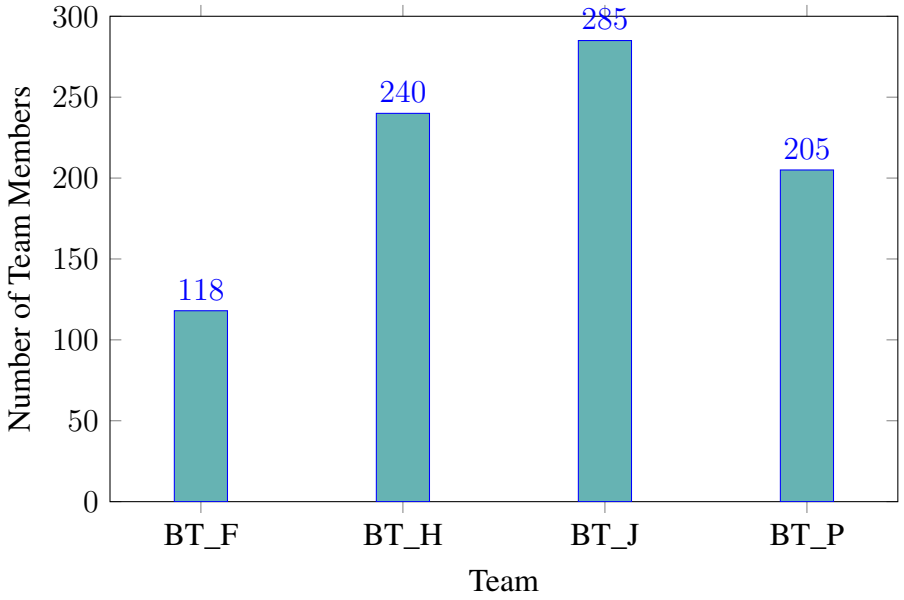

Figure 6. Team Sizes Reported in the FER

Observations:

O138: **BT_F**, with an estimated 118 members, was notably smaller compared to the other teams.

O139: **BT_H** was also relatively large at around 240, while **BT_P** came in at 205.

O140: **BT_J** was the largest team reported at approximately 285 personnel.

5. Evaluation of Blue Team TTPs

This section analyzes team-by-team the TTPs used by selected Blue Teams (Teams F, H, J, and P) during Locked Shields 2024, leveraging both qualitative insights gathered through semi-structured interviews and quantitative scoring metrics from the FER. The performance of each team is examined through metrics such as system availability, successful attack mitigations (computer systems, networks, web applications), report submissions (situation reports, threat intelligence, strategic communications and legal aspects), and overall usability of defended systems. Furthermore, the analysis considers each team's stated strategic objectives—specifically, whether their primary focus was achieving high exercise scores or emphasizing education and training outcomes. Observations made from qualitative interview data are correlated with quantitative performance indicators to reveal patterns and factors contributing to the effectiveness of TTP implementation. This integrated evaluation aims to identify and understand the relationships between team strategies, objectives, resource allocation, and measurable outcomes.

5.1 Blue Team F

This subsection analyzes the performance, strategies, and TTPs used by Blue Team F during the exercise.

5.1.1 Team Objectives and Structure

Blue Team F prioritized creating an educational environment where there are enough members of the team with experience from previous iterations of Locked Shields, from whom new members can learn, while also keeping the competitive element of the exercise alive. The 30-50% proportion of new members in the team seems to support that claim.[1]

Organizationally, the team was structured into subteams — including Network, Special Systems, Forensic, Pentest, Linux, Windows, Web, and SOC — each with a sub-lead reporting to a management team who lead the entire team. Members from each subteam were allocated to a specific zone.[2] This approach seems to shift low-level operational responsibilities to subteam leads rather than being handled by the central management team.

[1] Observation O16
[2] Observation O5

5.1.2 Key TTPs and Tools Used

A proactive containment strategy was used by sandboxing web applications and developing a proxy server that rewrote CDN connections dynamically. Each service was isolated in its own environment, possibly reducing the risk of lateral movement in the event of compromise.

During the interview with Team F, it was emphasized that they did not attempt to identify and patch every individual remote code execution (RCE) vulnerability. Instead, they recognized the futility of that approach in the context of Locked Shields, where attackers are assumed to possess complete knowledge of system vulnerabilities. The interviewee explained, "If we patch 19 out of 20 RCEs, and the Red Team knows about all 20, it's effectively the same as not patching any of them.".[3] It was also noted during the interview that the attackers were not able to go beyond the web applications themselves.

The core of the team's deployment and hardening strategy was built around Ansible, which was used to automate the rollout of configuration changes and apply sandboxing profiles across multiple systems. The automation was said to have generally worked although manual fixes were also necessary.

This strategy seems to be a trade-off. While the sandboxing approach successfully prevented attackers from moving laterally or causing full system compromise, it did not prevent web application exploitation. Their systems remained largely available throughout the exercise, and the final report reflected high availability and moderate usability scores in the web category, suggesting that the containment approach effectively limited the scope of the Red Team's impact.[4]

5.1.3 Communication and Coordination

Blue Team F used Discord for internal communication and coordination during the exercise. The team implemented a system within Discord to manage incident reporting and response. When a technical sub-team identified an incident, they could open a ticket in a designated reporting channel. A member of the reporting team would then be automatically assigned to the ticket and could ensure that the data being inputted to MISP or the UST is correct, by collecting technical details through follow-up questions and interviews with the corresponding BT member.

[3] Observation O32
[4] Observation O31

Feedback from the User Simulation team supports the effectiveness of this approach: Team F received a communication rating of **17.13**, indicating that their reporting processes were perceived as professional, responsive, and technically competent.[5] The same strategy likely contributed to their performance in the CTIREP category, where Team F received an above-average score.[6]

It was also noted during the interview that Team F had a high degree of operational flexibility by dynamically reallocating personnel throughout the exercise. Specifically, it was said that during day 2 of the exercise, when the team lead noticed that there was a shortage of personnel in the forensics department, they reallocated people to help out. The intervention appears to have come too late to recover fully in that area, as the final score in the forensics category remained below average.[7]

5.1.4 Strengths and Successes

One of Blue Team F's most successful strategies was their implementation of web application sandboxing combined with a custom proxy server for rewriting third-party CDN connections. From the interviewee's perspective, this was a clear validation of their containment-first strategy, as the systems remained stable and accessible even under constant attacks from the Red Team.[8]

Another notable strength was the team's adaptability in response to operational stress. When specific sub-teams, such as forensics or incident response, became overwhelmed, team leadership quickly reassigned available personnel to relieve bottlenecks and restore workflow continuity. While the effectiveness of these interventions varied across domains, it still reflects a high degree of internal coordination and trust.

5.1.5 Challenges Faced

Blue Team F encountered a number of challenges during the exercise that limited their overall performance. One of the key reflections from the interviewee was that the team had underestimated the amount of preparation required for Locked Shields.[9]

Internal communication also presented difficulties, the interviewee remarked that ensuring

[5] Observation O139
[6] Observation O130
[7] Observation O135
[8] Observation O31
[9] Observation O1

critical information reached the right individuals was not always successful.[10]

It was also highlighted that there was a challenge related to decision-making confidence among some team members. In several cases, individuals hesitated to take initiative and escalated decisions to more experienced colleagues. While this approach may have minimized immediate risk, it also created bottlenecks that slowed incident response. This issue might potentially be avoided by targeted pre-exercise training or earlier identification of individual skill gaps.[11]

5.1.6 Overall Assessment

Blue Team F demonstrated a strategically sound approach centred on containment, automation, and communication-driven coordination. Their use of sandboxing and a CDN-rewriting proxy for web services successfully maintained system availability, supporting the effectiveness of their containment-first strategy. Also, operational adaptability was another strength. Team leadership was able to dynamically reassign personnel to underperforming areas—such as forensics and incident response—based on emerging needs. While these reallocations were not always sufficient to recover lost ground, they reflect a mature, flexible internal structure. The team's communication model, built around Discord ticketing and a dedicated reporting group, also received positive evaluations, as reflected in their scores in the CTIREP category and UST feedback.

However, Team F's performance was constrained by several key challenges. Underestimation of required preparation time, communication friction, and uneven confidence levels among team members contributed to avoidable delays and inefficiencies.

In conclusion, Team F's TTP implementation was effective in achieving a resilient, available infrastructure, and structured external reporting, but also revealed areas where earlier preparation and internal process refinement could have produced stronger results across all categories.

5.2 Blue Team H

This subsection analyses the performance, strategies, and TTPs used by Blue Team H during the exercise.

[10] Observation O1
[11] Observation O1

5.2.1 Team Objectives and Structure

The goal for Blue Team H during Locked Shields 2024 was to create a learning environment where team members could gain exposure to realistic defensive operations and take those insights back to their home organizations.[12] This emphasis on learning was reinforced by the team's composition, as BT-H reported the highest proportion of new members. [13]

Blue Team H used a functional structure, which were created around areas such as Windows, Linux, Network, Web, and Forensics.[14]

5.2.2 Key TTPs and Tools Used

The core TTP for Team H was described as proactive reconnaissance during the familiarization phase. This included gathering data about the infrastructure, identifying potentially vulnerable systems, and compiling internal documentation to support the team once the exercise began.[15]

Based on this early reconnaissance, they created a structured plan for day one, focusing on hardening and patching endpoints and setting up monitoring. This approach helped streamline team coordination and reduce reactive decision-making during the beginning of Day 1.[16] This emphasis on structured preparation is consistent with BT-H's strong performance in communication and documentation, as the team achieved notably high CTIREP scores. [17]

Automation tools were leveraged to achieve baseline hardening and deployment of agents on endpoints.[18]

5.2.3 Communication and Coordination

A reporting procedure was also implemented by using an internal ticketing system where the incident coordinators were used to help communicate with the UST and provide information to the reporting sub team.[19]

[12]Observation O98, O99, O100
[13]Observation O17
[14]Observation O5
[15]Observation O26
[16]Observation O26
[17]Observation O128
[18]Observation O61
[19]Observation O26

Subteam leads managed tasks and imposed a 5-15 minute time limit for the team members before they had to escalate the issue to senior personnel.[20]

5.2.4 Strengths and Successes

The goal for them was to get the Locked Shields experience to as many people as possible. This objective was reflected in the team's composition, which prioritized broad participation and ensured that a large share of members were able to gain firsthand experience of the exercise. [21]

They also had a high-performing legal subteam, which is reflected in the high score they received in the legal category. [22]

5.2.5 Challenges Faced

Patching web applications was a challenge for them, as they described that the web applications that get developed for the Locked Shields exercise are incredibly complex in nature and therefore hard to secure.[23] This also appears to be reflected in the FER, where the team received a large penalty score for the undefended Red Team's web-based attacks.[24]

It was also noted during the interview that a high variance in latency to the game net environment caused some technical difficulties for them.[25]

5.2.6 Overall Assessment

Overall, Blue Team H's TTP implementation demonstrated thoughtful preparation, organizational structure, and a commitment to the team's learning-oriented mission. While technical performance in web security and defending computer systems may have suffered as a result of that prioritization, the team's approach succeeded in fulfilling its stated educational objectives.[26]

[20]Observation O37
[21]Observation O98, O99, O108
[22]Observation O133
[23]Observation O127
[24]Observation O134
[25]Observation O2
[26]Observation 134

5.3 Blue Team J

This subsection analyses the performance, strategies, and TTPs used by Team J during the exercise.

5.3.1 Team Objectives and Structure

Blue Team J's main goal was to win Locked Shields 2024 exercise.[27] In addition to this main goal, the team also aimed to strengthen cooperation with allied nations.[28] It was also noted that a key outcome was to train military and leadership roles and for the personnel to have the opportunity to learn and develop their skills.[29] Their primary objective seems to be supported by the final total score reported in the FER.[30]

The team used a functional structure, organizing members into specialized sub teams based on technical domains such as Windows, Linux, Web, and Forensics. Each sub team had their own team lead who managed the low-level operational tasks. [31] A notable remark is also that Blue Team J had an "ambulance" sub team dedicated to providing quick responses to escalated incidents.[32]

5.3.2 Key TTPs and Tools Used

Blue Team J demonstrated a comprehensive implementation of defensive TTPs across the preventive, detective, and responsive categories. These practices reflect a high degree of pre-exercise preparation, tactical foresight, and integration of tooling with operational workflows—consistent with the team's stated objective of winning the exercise.

In the preventive domain, each sub team came up with their own detailed plan confined to their technical scope, supported by a centralized code and tooling repository and an introductory guide for team members.[33] A particularly notable initiative was the creation of a detailed 30+30-minute execution plan for the first hour of the exercise, outlining detailed technical plans coordinated among sub teams that were separated into different phases depending on the priority of actions. These plans were enriched by proactive

[27] Observation O101
[28] Observation O104
[29] Observation O102, O103
[30] Observation 137
[31] Observation O38
[32] Observation O38
[33] Observation O27

reconnaissance efforts during the familiarization phase.[34]

Blue Team J also focused heavily on automation by scraping web applications and hardening them through application containerization and the deployment of a web application firewall, indicating a layered defence approach.[35] This seems to be supported by the score in the FER, where Blue Team J received the lowest score among the 4 Blue Teams.[36]

On the detective side, endpoint visibility was enhanced through the deployment of both monitoring and EDR agents, improving telemetry and threat detection at the host level.[37] Custom tooling was also used to enrich and correlate events gathered from the monitoring agents.[38]

The team's response workflows were methodically structured. An internal ticketing system enhanced by Discord bots was used to facilitate the triage of incidents and tickets and assign them to the appropriate technical sub teams.[39] In addition, a shared network map and role-responsibility matrix improved situational awareness and reduced friction during coordinated incident response efforts.[40]

These TTPs indicate a technically mature team operating under a cohesive strategy. The emphasis on automation, pre-task allocation, and internal coordination infrastructure suggests that Team J prepared extensively. [41]

5.3.3 Communication and Coordination

Blue Team J had a hybrid communication model, with part of the team operating on-site and the rest collaborating remotely via digital channels.[42] Procedures were made for communication workflows to ensure escalation and accuracy.[43]

When a sub team was unable to resolve an issue autonomously, the incident would be escalated to a specialized "ambulance" team—composed of highly skilled responders tasked with resolving complex problems.[44] This structure, in theory, indicates that incidents

[34]Observation O27
[35]Observation O70
[36]Observation O137
[37]Observation O27
[38]Observation O62
[39]Observation O64
[40]Observation O27
[41]Observation O59
[42]Observation O43
[43]Observation O27
[44]Observation O38

were handled efficiently across the entire scope of the exercise, which also seems to be supported by the final total score.[45]

A dedicated help desk was established to communicate with the UST, streamlining the handling of service tickets and reports. This seems to be supported by the high usability score.[46]

5.3.4 Strengths and Successes

It was noted that a outbound connection firewall hardening prevented malware from reaching C2 servers, showing that some preventive actions worked as planned.

Blue Team J showed that they could adapt their strategies on the fly by noticing during the exercise that the procedure for firewall rule requesting was becoming congested and needed resolving. A change in that procedure was implemented to eradicate the issue.[47]

While Blue Team J did not achieve first place as they intended they still got a solid spot in the top 3.

5.3.5 Challenges Faced

A notable failure occurred during the first phase of the exercise, where a hardening script was never executed due to a breakdown in communication. As a result, Phase 1 user-side attacks were marked as successful in the scoring and monitoring system, contributing to a penalty in that category.[48]

Some automation scripts failed to run due to the LS environment being different from las year, but this was discovered through doing comprehensive full team rehearsals on Day 0, which allowed them to correct the mistake.[49] Also some monitoring agents failed to work due to the performance constraints on the game net VMs.[50]

[45] Observation O132
[46] Observation O138
[47] Observation O84
[48] Observation O34
[49] Observation O76
[50] Observation O78

5.3.6 Overall Assessment

Blue Team J demonstrated a strategically mature and well-prepared approach to Locked Shields 2024, clearly aligned with their primary objective of winning the exercise. Their functional team structure, layered with a high-skill ambulance team, allowed for clear task allocation, vertical escalation, and coordinated response across technical domains.

However, despite their extensive planning, Blue Team J was not immune to critical setbacks. A miscommunication during the early phase of the exercise resulted in a key automation script not being executed, which contributed to a penalty in user-side attack categories.

In summary, Blue Team J successfully deployed a comprehensive defensive strategy through disciplined planning, automation, and internal coordination. Their performance reflects a high level of Blue Team maturity, and while they fell short of winning the competition, their placement in the top three indicates that their methods were broadly effective.

5.4 Blue Team P

This subsection analyzes the performance, strategies, and TTPs used by Team P during the exercise.

5.4.1 Team Objectives and Structure

Blue Team P went into Locked Shields 2024 with the clear objective of winning the exercise.[51]

The team adopted a functional structure, dividing responsibilities across specialized sub-teams such as Windows, Linux, Web, and Forensics. A sub team was dedicated that was responsible for overseeing the active tasks and incidents.[52] In addition to this, Blue Team P introduced a coordination layer through the use of "segment owners," whose primary function was to oversee intra-sub team communications.

[51] Observation O105
[52] Observation O39

5.4.2 Key TTPs and Tools Used

Blue Team P adopted a highly structured and automation-driven approach to defensive operations during Locked Shields 2024. Their preventive strategies included operating system and firewall hardening, supported by a specialized support team tasked with developing integrations and automation scripts across the team's infrastructure.[53] This effort was reinforced by the distribution of an introductory guide to all team members and the formulation of a detailed 30-minute plan for the initial phase of the exercise.[54]

Blue Team P aimed to automate as many defensive tasks as feasible. The first 30 minutes of the exercise was driven by automations, while the remaining phases relied more on manual tasks. This strategy reflects a balanced approach to automation for rapid initial response while maintaining flexibility to act manually as the scenario develops.[55]

Detection capabilities were strengthened through the deployment of EDR agents and endpoint monitoring tools. These provided host-level visibility and contributed to a proactive threat-hunting posture. Blue Team P implemented custom tooling to support environment-specific automation workflows, suggesting a high level of technical adaptation.[56]

On the responsive side, Blue Team P formalized its processes using wiki-based documentation, clearly defined internal reporting procedures, and escalation pathways. These measures aimed to ensure that critical incidents were addressed efficiently and consistently across subteams.[57]

5.4.3 Communication and Coordination

Communication for Blue Team P was handled by having the team participate in the exercise in an on-site manner, it was also complemented by chat servers for information sharing.[58]

5.4.4 Strengths and Successes

Blue Team P reported being very prepared for the exercise[59], due to hosting two large internal exercises prior to the execution of LS. One of these exercises was purely for

[53] Observation O72
[54] Observation O79
[55] Observation O73, O79
[56] Observation O72
[57] Observation O28
[58] Observation O45
[59] Observation O60

the technical team, while the other one was a full-team exercise integrating together the non-technical and technical parts of the team.[60]

5.4.5 Challenges Faced

Blue Team P did not report experiencing major operational challenges during the exercise. However, they noted difficulties regarding the reporting aspect of the competition.[61] Specifically, they expressed concern that, due to the subjective nature of the reporting evaluation, the feedback that was received during the exercise did not consistently offer clear or actionable guidance on how to improve their reporting procedures. Moreover, it was explained that the reporting expectations within the exercise differed from real-world operational practices.

5.4.6 Overall Assessment

Blue Team P demonstrated a highly structured and prepared approach during Locked Shields 2024. Their functional organization, emphasis on automation, and strong detection capabilities contributed to a resilient and coordinated defence. Preparation through prior internal exercises further strengthened team cohesion and operational readiness.

Although no major operational challenges were reported, the team encountered some difficulties with the subjective nature of reporting feedback, highlighting a gap between real-world practices and exercise-specific expectations. Nevertheless, Blue Team P adapted effectively and maintained a consistently high level of performance throughout the competition.

[60] Observation O60
[61] Observation O4

6. Discussion

This chapter interprets the findings presented in the evaluation chapter and contextualizes them within broader cybersecurity defence principles. It aims to explore patterns and contrasts that emerged across teams and reflect on their implications for real-world Blue Team operations. By reflecting on both the successes and limitations observed in the evaluated teams, the chapter seeks to contribute to a deeper understanding of what constitutes effective defensive practice in live-fire cyber exercises like Locked Shields—and what lessons can be applied beyond them.

6.1 Interpretation of Results

One of the clearest patterns to emerge from the evaluation was the influence of each team's strategic objectives and structural choices on their performance and coordination during the exercise. While all teams used a functional subteam structure to some extent the final structure differed in their implementation.

Teams J and P entered Locked Shields with a clear and singular goal: to win the exercise. This competitive focus was reflected in their use of well-defined functional teams and in the case of Team P, an added coordination mechanism via "segment owners," who oversaw communication and task flow across subteams. Team J on the otherhand opted to include a special "ambulance" team for handling issues that could not be resolved by the subteams. These team structures show that a well coordinated team was put together and suggest a higher degree of organizational maturity, likely aimed at optimizing response times and reducing friction in high-pressure scenarios.

Teams F and H, in contrast, approached the exercise with a more educational objective. While not identical in team structure their results — particularly the difficulty in securing web services — may reflect the challenges of balancing the educational aspect and technical depth.

6.2 Answering the Research Questions

This subsection summarises how the findings of this study address the research questions defined in Chapter 1. Through the integration of qualitative and quantitative findings, the study directly addresses both RQ1 and RQ2.

RQ1: How can the effectiveness of Blue Team TTPs during the Locked Shields 2024 exercise be systematically evaluated and correlated with specific defence outcomes?

The results of this thesis demonstrate that the effectiveness of Blue Team TTPs can be systematically evaluated through a mixed-methods approach. Three elements were essential in answering RQ1:

1. **Qualitative insights.** Semi-structured interviews provided detailed descriptions of planning, execution, and adaptation of TTPs. The structured categorisation of interview questions enabled systematic comparison across teams in areas such as preparedness, automation usage, communication pipelines, team composition, and adaptability.
2. **Quantitative performance data.** Metrics from the Final Exercise Report provided measurable indicators of defensive effectiveness. These data points allowed the study to link TTPs to specific defence outcomes.
3. **Integration of datasets.** By cross-referencing qualitative themes with quantitative results, the study was able to identify correlations between TTP implementation and defence outcomes.

Together, these components form a reproducible framework for evaluating Blue Team TTPs in live-fire exercises, thereby directly answering RQ1.

RQ2: What strategies can be recommended to improve Blue Team coordination and decision-making during live-fire exercises, based on the evaluation of TTP performance?

The evaluation revealed several recurring patterns across teams that directly inform strategies for improving coordination and decision-making. The following recommendations are grounded in observed correlations between specific TTPs and team performance:

1. **Establish clear internal communication pipelines.** Teams that implemented structured communication workflows (e.g., internal ticketing systems, escalation mechanisms, “ambulance teams”) responded more efficiently to incidents and achieved higher reporting-related scores.
2. **Implement escalation structures and role clarity.** Mechanisms such as segment owners (BT_P) or specialised response subteams (BT_J) seemed to reduce bottle-

necks and improve incident response.

3. **Strengthen pre-exercise rehearsals and testing.** Teams that conducted Day 0 rehearsals and internal simulations (BT_J, BT_P) detected automation failures early and demonstrated smoother coordination during the exercise.
4. **Reinforce IT operational foundations.** As emphasised by BT_P, strong cyber defence depends on stable IT operations. Weak underlying operational practices increase defensive workload and reduces decision-making efficiency.
5. **Support decision-making confidence through training.** As seen in BT_F, hesitation to make decisions slowed response times. Targeted training and responsibility clarification strengthen decision-making under pressure.

These findings provide empirically grounded strategies for enhancing Blue Team performance in future live-fire exercises and real-world cyber defence operations, directly addressing RQ2.

6.3 Implications for Blue Teams

The findings highlight several important implications for Blue Team organization and operational practices in both exercise and real-world settings.

Firstly, clear and well-structured internal communication pipelines seemed to improve coordination and incident response capabilities. Teams that established internal escalation protocols (e.g., ambulance teams, segment owners) were better equipped to handle high incident loads without overwhelming individual members.

Secondly, while automation remains a critical enabler of early defensive operations, excessive reliance without accounting for real-time validation can introduce potential issues. Successful teams coupled automation with manual fallback procedures, ensuring continuity when scripted actions failed.

Thirdly, the results suggest that balancing between technical expertise and operational leadership is vital. Teams that separated purely technical tasks from coordination and management tasks — particularly in high-pressure environments — were able to sustain efficiency longer.

Finally, the divergence between real-world operational procedures and exercise scoring expectations remains a challenge. Blue Teams must be prepared to adapt their practices to the nuances of the exercise environment while maintaining operational realism.

6.4 Recommendations for Improvement

Based on the evaluation and observed practices, several recommendations can be made:

- **Enhance early planning and rehearsal cycles:** Teams that conducted structured Day 0 rehearsals and internal exercises performed better, particularly in detecting automation gaps and optimizing coordination workflows. Pre-exercise full-team simulations and rehearsals should be prioritized.
- **Formalize escalation and ticketing processes:** Introducing internal escalation mechanisms (e.g., ambulance teams, segment owners) seems to help with problem resolution speed and reducing decision-making bottlenecks during critical incidents.
- **Balance automation with manual readiness:** Blue Teams should ensure automation systems are thoroughly tested and validated, with manual fallback plans prepared for key tasks like threat hunting, endpoint recovery, and agent deployment.
- **Focus on operational IT foundations:** As highlighted by Team P, technical cyber defence efforts must be built on top of robust IT operations capabilities. Defending poorly maintained or unstable systems exacerbates vulnerabilities, regardless of cyber-specific expertise. [1]

6.5 Limitations and Challenges

Several limitations must be acknowledged when interpreting the findings of this study.

Firstly, the sample size was restricted to four Blue Teams due to limited participation in the interview phase. While the selected teams represent a meaningful cross-section of approaches, the results may not fully generalize across all participants of Locked Shields 2024.

Secondly, the qualitative data is based on self-reported information collected approximately nine months after the exercise, introducing potential recall bias.

Thirdly, the exercise scoring system itself introduces inherent subjectivity, particularly in categories like strategic communications and reporting, which complicates direct comparisons across teams.

Finally, the study's focus on a single edition of the Locked Shields exercise, while beneficial for depth, limits temporal generalizability. Trends observed may evolve as the Locked

[1] Observation O119

Shields format and threat landscape change over time.

Despite these limitations, the integration of qualitative and quantitative data provides a foundation for evaluating Blue Team TTPs and generating actionable insights for improving defensive practices in future exercises and real-world operations.

In conclusion, the evaluation of Blue Team TTPs at Locked Shields 2024 reveals both consistent strategies and unique adaptations that contributed to overall team performance. While organizational structure, communication, and automation emerged as key success factors, challenges related to exercise-specific dynamics and resource constraints remain evident. These findings provide a foundation for the final chapter, where the main contributions of this study are summarized and future research directions are proposed.

7. Summary

The primary objective of this thesis was to evaluate the effectiveness of Blue Team tactics, techniques, and procedures (TTPs) during the Locked Shields 2024 exercise. By integrating qualitative insights gathered from interviews with key Blue Team members and quantitative performance data from the Final Exercise Report (FER), this study provides a detailed understanding of which strategies contributed to successful cyber defence under live-fire conditions.

This chapter summarizes the main contributions of the study and outlines recommendations for future research.

7.1 Summary of Contributions

This research makes several key contributions to the field of cyber defence evaluation:

- **Systematic Evaluation of Blue Team TTPs:** This study offers a structured evaluation framework demonstrating how Blue Team TTPs can be systematically analysed through semi-structured interviews. It emphasizes the methodological strength of gathering qualitative insights from team members and correlating these findings with quantitative performance metrics obtained from the Locked Shields 2024 Final Exercise Report.
- **Analysis of Team Structures and Strategic Objectives:** By comparing teams with different strategic goals—whether educational or competitive—the study reveals how organizational structure and internal priorities influence operational effectiveness during high-pressure exercises.
- **Identification of Success Factors and Challenges:** Common success factors such as early preparation, escalation mechanisms (e.g., ambulance teams, segment owners), and hybrid communication models were identified. Conversely, recurring challenges such as adapting to subjective scoring systems, automation failures, and resource constraints were also analysed.

Collectively, these contributions advance understanding of how defensive teams operate in dynamic, adversarial scenarios and provide a foundation for improving future cybersecurity exercises and real-world incident response planning.

7.2 Future Research Directions

While this study provides valuable insights, it also opens several avenues for further research:

- **Broader Sample Size:** Future studies could benefit from including a larger number of teams across multiple years of Locked Shields to validate and refine the observed patterns over time.
- **Impact of Emerging Technologies:** The increasing integration of artificial intelligence and machine learning in both offensive and defensive cyber operations warrants investigation. Studying how AI-enabled tools affect Blue Team workflows would be a valuable extension.

Through these avenues, future research can continue to refine and strengthen the effectiveness of Blue Team practices against evolving cyber threats.

References

[1] Li Yuchong and Liu Qinghui. "A comprehensive review study of cyber-attacks and cyber security; Emerging trends and recent developments". In: *Energy Reports* 7 (2021), pp. 8176–8186. ISSN: 2352-4847. DOI: https://doi.org/10.1016/j.egyr.2021.08.126. URL: https://www.sciencedirect.com/science/article/pii/S2352484721007289.

[2] CCDCOE. *Locked Shields*. [Accessed: 12-02-2025]. URL: https://ccdcoe.org/locked-shields/.

[3] CCDCOE. *Locked Shields*. [Accessed: 09-11-2025]. URL: https://ccdcoe.org/news/2024/locked-shields-2024-sets-the-stage-for-advancing-global-cyber-defence/?utm_source=chatgpt.com.

[4] Max Smeets. "The Role of Military Cyber Exercises: A Case Study of Locked Shields". In: *2022 14th International Conference on Cyber Conflict: Keep Moving! (CyCon)*. Vol. 700. IEEE, May 2022, pp. 9–25. DOI: 10.23919/cycon55549.2022.9811018.

[5] Manuel Kern et al. *Towards Improving IDS Using CTF Events*. 2025. arXiv: 2501.11685 [cs.CR]. URL: https://arxiv.org/abs/2501.11685.

[6] Jan Vykopal et al. "Research and Practice of Delivering Tabletop Exercises". In: *Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1*. ITiCSE 2024. ACM, July 2024, pp. 220–226. DOI: 10.1145/3649217.3653642.

[7] Sten Mäses. "Evaluating Cybersecurity-Related Competences through Simulation Exercises". PhD thesis. TalTech, 2020. DOI: 10.23658/taltech.49/2020.

[8] Kaie Maennel. "Advancing Cybersecurity Education through Learning Analytics". PhD thesis. TalTech, 2021. DOI: 10.23658/taltech.29/2021.

[9] Kathryn Parsons et al. "The Human Aspects of Information Security Questionnaire (HAIS-Q): Two further validation studies". In: *Computers Security* 66 (2017), pp. 40–51. ISSN: 0167-4048. DOI: https://doi.org/10.1016/j.cose.2017.01.004. URL: https://www.sciencedirect.com/science/article/pii/S0167404817300081.

[10] Margus Ernits et al. "From Simple Scoring Towards a Meaningful Interpretation of Learning in Cybersecurity Exercises". In: Mar. 2020. DOI: 10.34190/ICCWS.20.046.

[11] A. Nicholson et al. “SCADA security in the light of Cyber-Warfare”. In: *Computers Security* 31.4 (2012), pp. 418–436. ISSN: 0167-4048. DOI: https://doi.org/10.1016/j.cose.2012.02.009. URL: https://www.sciencedirect.com/science/article/pii/S0167404812000429.

[12] Daniel Schatz, Rabih Bashroush, and Julie Wall. “Towards a More Representative Definition of Cyber Security”. In: *Journal of Digital Forensics, Security and Law* 12 (June 2017), p. 53. DOI: 10.15394/jdfsl.2017.1476.

[13] CCDCOE. *Locked Shields 2024 Final Exercise Report*.

[14] Claire La Fleur et al. “Team performance in a series of regional and national US cybersecurity defense competitions: Generalizable effects of training and functional role specialization”. In: *Computers & Security* 104 (May 2021), p. 102229. ISSN: 0167-4048. DOI: 10.1016/j.cose.2021.102229. URL: https://www.sciencedirect.com/science/article/pii/S0167404821000535.

[15] Magdalena Granåsen and Dennis Andersson. “Measuring team effectiveness in cyber-defense exercises: a cross-disciplinary case study”. In: *Cognition, Technology & Work* 18.1 (Feb. 2015), pp. 121–143. ISSN: 1435-5566. DOI: 10.1007/s10111-015-0350-2.

[16] Andrea Skytterholm and Guro Hotvedt. “Criteria for Realistic and Expedient Scenarios for Tabletop Exercises on Cyber Attacks Against Industrial Control Systems in the Petroleum Industry”. In: *Proceedings of the International Conference on Cybersecurity, Situational Awareness and Social Media*. Springer Nature Singapore, Mar. 2023, pp. 39–54. ISBN: 9789811964145. DOI: 10.1007/978-981-19-6414-5_3.

[17] Yongjoo Shin et al. “A Study on Designing Cyber Training and Cyber Range to Effectively Respond to Cyber Threats”. In: *Electronics* 13.19 (2024). ISSN: 2079-9292. DOI: 10.3390/electronics13193867. URL: https://www.mdpi.com/2079-9292/13/19/3867.

[18] Hannes Holm and Teodor Sommestad. “Realistic and balanced automated threat emulation”. In: *Computers & Security* 151 (2025), p. 104351. ISSN: 0167-4048. URL: https://www.sciencedirect.com/science/article/pii/S0167404825000409.

[19] Marko Arik, Adrian Venables, and Rain Ottis. “The Optimal Organisational Structure for Cyber Operations Based on Exercise Lessons.” In: *European Conference on Cyber Warfare and Security* 23 (June 2024), pp. 37–48. DOI: 10.34190/eccws.23.1.2244.

[20] Claire La Fleur, Blaine Hoffman, and C. Gibson. “Team Performance in a Series of Regional and National US Cybersecurity Defense Competitions: Generalizable Effects of Training and Functional Role Specialization”. In: *Computers & Security* 104 (Feb. 2021), p. 102229. DOI: 10.1016/j.cose.2021.102229.

[21] Norbou Buchler et al. “Cyber Teaming and Role Specialization in a Cyber Security Defense Competition”. In: *Frontiers in Psychology* 9 (2018). ISSN: 1664-1078. DOI: 10.3389/fpsyg.2018.02133. URL: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2018.02133.

[22] Sten Mäses, Kaie Maennel, and Agnė Brilingaitė. “Trends and challenges for balanced scoring in cybersecurity exercises: A case study on the example of Locked Shields”. In: *Frontiers in Education* 7 (Sept. 2022). ISSN: 2504-284X. DOI: 10.3389/feduc.2022.958405. URL: https://www.researchgate.net/publication/363672892_Trends_and_challenges_for_balanced_scoring_in_cybersecurity_exercises_A_case_study_on_the_example_of_Locked_Shields.

[23] Federica Bianchi, Enrico Bassetti, and Angelo Spognardi. “Scalable and automated Evaluation of Blue Team cyber posture in Cyber Ranges”. In: *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing*. ACM, Apr. 2024, pp. 1539–1541. DOI: 10.1145/3605098.3636154. URL: http://dx.doi.org/10.1145/3605098.3636154.

[24] Kunal Sehgal and Nikolaos Thymianis. *Cybersecurity Blue Team Strategies*. Packt Publishing, 2023. ISBN: ISBN: 978-1-80107-247-2.

[25] Allard Dijk et al. “Next Steps in Cyber Blue Team Automation—Leveraging the Power of LLMs”. In: May 2025, pp. 209–226. DOI: 10.23919/CyCon65856.2025.11103720.

[26] Splunk. *What Are TTPs? Tactics, Techniques & Procedures Explained*. [Accessed: 14-02-2025]. URL: https://www.splunk.com/en_us/blog/learn/ttp-tactics-techniques-procedures.html.

[27] Fernando Maymi et al. “Towards a definition of cyberspace tactics, techniques and procedures”. In: *2017 IEEE International Conference on Big Data (Big Data)*. IEEE, Dec. 2017. ISBN: ISBN: 978-1-5386-2715-0. DOI: 10.1109/bigdata.2017.8258514.

Appendix 1 – Non-Exclusive License for Reproduction and Publication of a Graduation Thesis[1]

I Kristo Tammsoo

1. Grant Tallinn University of Technology free licence (non-exclusive licence) for my thesis "Evaluating Blue Team TTPs against cyber threats on the example of Locked Shields 2024", supervised by Rain Ottis and Bernt Åkesson
   1.1. to be reproduced for the purposes of preservation and electronic publication of the graduation thesis, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright;
   1.2. to be published via the web of Tallinn University of Technology, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright.
2. I am aware that the author also retains the rights specified in clause 1 of the non-exclusive licence.
3. I confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights, the rights arising from the Personal Data Protection Act or rights arising from other legislation.

January 9, 2026

[1]The non-exclusive licence is not valid during the validity of access restriction indicated in the student's application for restriction on access to the graduation thesis that has been signed by the school's dean, except in case of the university's right to reproduce the thesis for preservation purposes only. If a graduation thesis is based on the joint creative activity of two or more persons and the co-author(s) has/have not granted, by the set deadline, the student defending his/her graduation thesis consent to reproduce and publish the graduation thesis in compliance with clauses 1.1 and 1.2 of the non-exclusive licence, the non-exclusive license shall not be valid for the period.

Appendix 2 - Semi-Structured Interview Questions

General Questions

1. Can you briefly describe your role in the exercise?
2. What is your daily job/duty if you are allowed to talk about it?
3. How many times have you participated Locked Shields?
4. In one sentence describe how was your team's experience in Locked Shields 2024?
5. What were the most significant challenges your team faced during the exercise?

Team Composition and Resources

6. How was your team structured?
7. What was the level of expertise in administration, forensics, incident response, automization, threat analysis, and pentesting within your team?
8. What was the proportion of new vs veteran individuals in the team?
9. Did the size of your team impact your ability to execute defensive strategies effectively?

Gathering Information About TTPs

10. Can you describe your planned TTPs?
11. Were there specific defensive actions that had a measurable impact on system uptime or other performance metrics? Why do you think so, how did you measure?
12. How did you allocate resources such as time and team members to address multiple attack vectors?
13. How did your team communicate and coordinate tasks during high-pressure moments?
14. Were there any TTPs your team had planned that turned out to be ineffective?
15. Were you able to test and prepare the TTPs prior to the exercise?
16. How would you rate your team's preparedness for the exercise?

Automation and Tools

17. Describe the automated tools or processes that your team used.

18. How much time did you spend on automation vs manual tasks?
19. Were there scenarios where automation failed or where manual intervention was necessary?

Adaptability

20. Were there moments where your team had to abandon one TTP in favor of another? What prompted this shift?
21. Can you share an example where improvisation played a key role in your defensive actions?

Evaluation and Lessons Learned

22. Did you have any internal goals, and how did you plan to achieve them? (If yes, then follow up.)
23. Do you think they were effective in achieving those goals? Why do you think so?
24. What were the most significant takeaways from the exercise that you would apply in future scenarios?
25. Did you have an internal lessons learned process?
26. In hindsight, were there areas where additional preparation or resources could have improved the performance of your team?

Appendix 3 - Interview Categorization Table

Question Code	Variable Name	Value	Definition
Q1	Q1_Role	Technical role	
Q1	Q1_Role	Strategy & Leadership role	
Q2	Q2_Category	Cybersecurity role	
Q2	Q2_Category	IT & Technical role	
Q2	Q2_Category	Managerial role	
Q2	Q2_Category	Academia/Research role	
Q2	Q2_Category	Military/Defense role	
Q3	Q3_Count	Numeric	
Q3	Q3_Category	0	
Q3	Q3_Category	1	
Q3	Q3_Category	2-4	
Q3	Q3_Category	4+	
Q4	Q4_Experience_-Sentence	Free text	
Q4	Q4_Experience_-Sentiment	Positive	
Q4	Q4_Experience_-Sentiment	Neutral	
Q4	Q4_Experience_-Sentiment	Negative	
Q5	Q5_Challenge_-Category	Communication Issues	
Q5	Q5_Challenge_-Category	Technical Difficulties	
Q5	Q5_Challenge_-Category	Resource Constraints	
Q5	Q5_Challenge_-Category	Coordination/Logistics	
Q5	Q5_Challenge_-Category	Time Management	

Question Code	Variable Name	Value	Definition
Q5	Q5_Challenge_-Category	Training/Skill Gaps	
Q6	Q6_Team_Structure_Type	Functional	
Q6	Q6_Team_Structure_Type	Zone based	
Q7	Q7_Expertise_-Admin	Numeric (1–5)	
Q7	Q7_Expertise_-Forensics	Numeric (1–5)	
Q7	Q7_Expertise_Incident_Response	Numeric (1–5)	
Q7	Q7_Expertise_-Automization	Numeric (1–5)	
Q7	Q7_Expertise_-Threat_Analysis	Numeric (1–5)	
Q7	Q7_Expertise_-Pentesting	Numeric (1–5)	
Q7	Q7_Team_Composition_By_Expertise	Low	
Q7	Q7_Team_Composition_By_Expertise	Medium	
Q7	Q7_Team_Composition_By_Expertise	High	
Q8	Q8_Percentage_-New	Numeric (0-100)	
Q8	Q8_Composition_Category	Mostly New	if >60% new
Q8	Q8_Composition_Category	Balanced	if ∼40–60% new
Q8	Q8_Composition_Category	Mostly Veteran	if <40% new

Question Code	Variable Name	Value	Definition
Q9	Q9_Optimal_-Team_Size_Perception	Understaffed	
Q9	Q9_Optimal_-Team_Size_Perception	Optimal	
Q9	Q9_Optimal_-Team_Size_Perception	Overstaffed	
Q10	Q10_TTP_Category_Preventive		Examples: Firewall enhancements, patch management, access controls, system hardening, secure configurations.
Q10	Q10_TTP_Category_Detective		Examples: Intrusion detection systems, log monitoring, SIEM alerts, anomaly detection tools, network traffic analysis.
Q10	Q10_TTP_Category_Responsive		Examples: Incident response playbooks, automated response protocols, escalation procedures, crisis communication plans, forensic analysis processes.
Q11	Q11_Measurable_Impact_-Text		
Q12	Q12_Resource_-Allocation_Text		
Q13	Q13_Communication_Channel_-Text		
Q14	Q14_TTP_Ineffective_Text		
Q15	Q15_Testing_-Text		
Q16	Q16_Preparedness_Rating	Numeric (1–5)	

Question Code	Variable Name	Value	Definition
Q17	Q17_Automation_Tools		
Q18	Q18_Automation_Text		
Q19	Q19_Failure_Text		
Q20	Q20_TTP_Shift_Text		
Q21	Q21_Improvisation_Text		
Q22	Q22_Internal_Goals_Text		
Q23	Q23_Goals_Achieved_Text		
Q24	Q24_Takeaways_Category_Text		
Q25	Q25_LessonsLearned_Text		
Q26	Q26_Hindsight	Yes	
Q26	Q26_Hindsight	No	
Q26	Q26_Resource_Preparation_Category_Text		