Cyber-pi: Intelligent cyberthreat detection and supervised response

Alexandros Papanikolaou[1,*], Christos Ilioudis[2] and Vasilis Katos[3]

[1]*Innovative Secure Technologies P.C., Thessaloniki, Greece*
[2]*International Hellenic University, Thessaloniki, Greece*
[3]*Bournemouth University, Fern Barrow, Poole, UK*

Abstract

Integration of cyber incident management systems comes with a series of challenges on the organisational, technical and human dimension. In this paper we introduce Cyber-pi, a reference architecture for integrated cyber threat detection and response. This architecture is used to facilitate the study of the human aspects and showcases the interplay between the human and automated operator; these two dimensions are represented by the SIEM interface and the self-healing component of Cyber-pi respectively.

Keywords

integrated incident management, self-healing, human in the loop

1. Introduction and motivation

Cyber threats, following the trajectory of technological advances, become increasingly sophisticated. This trend highlights the need to revisit threat detection and response [1, 2]. According to the current state of the art and the heterogeneity and complexity of modern ICT infrastructures, a holistic and integrated approach in handling security incidents seems to be among the solutions that are of adequate effectiveness [3]. An integrated incident management system, in particular, can provide situational awareness across its constituency, including the organisation's devices and assets, applications, and business operations.

When considering integrated incident response systems, their effectiveness is limited by a number of factors that that may impair the operation and limit the added value of the incident response solution. The recent explosion and adoption of AI in various application domains provides a first glimpse of the benefits of AI facilitated workflows. It has also showed the dangers and risks when the integration of AI is performed in haste and not thoughtfully planned [4]. In the past few months there have been a number of online articles published describing how AI chatbots such as ChatGPT can be used in cybersecurity, with the majority relating to practical penetration testing activities, see for example [5]. Although the added value of AI-facilitated pentesting for both red and blue teams can be easily recognised, other areas such

Research Projects Track @ RCIS 2023: The 17th International Conference on Research Challenges in Information Science, May 23–26, 2023, Corfu, Greece

a.papanikolaou@innosec.gr (A. Papanikolaou); iliou@ihu.gr (C. Ilioudis); vkatos@bournemouth.ac.uk (V. Katos)
0000-0002-0251-0990 (A. Papanikolaou); 0000-0002-8084-4339 (C. Ilioudis); 0000-0001-6132-3004 (V. Katos)

as cyber incident detection and response are not yet adequately researched and explored. In fact, at the current state of play, we raise concerns that AI-enabled incident handling may cause more problems than remedies and could have potentially catastrophic consequences for the SOC team and the organisation as a whole.

This paper leverages a proposed cyberthreat detection and response architecture to pinpoint the components which will merit from further research on the interplay between AI-enabled incident handling and critical human factors. We propose an approach that considers factors of an attack (complexity, uncertainty, phase) and show how this can be combined with a human operator assessment framework (NASA TLX).

2. The CTI and response architecture

Figure 1 depicts the overall proposed architecture for integrated cyber incident detection and response. The Cyber-pi platform integrates established and popular technologies and standards for cyber threat intelligence data collection and sharing, operational security, asset management and visualisation. The components of interest to this paper are those of visualisation and self-healing. Appropriate dashboards built on the well-known OpenSearch stack[1] represent threat-related information, associated to the monitored assets. The information is further enriched with external CTI data feeds. The self-healing component offers the automated, intelligent cyberthreat detection services. The self-healing module in turn consists of the OpenC2 communication language to allow the remote execution of cyber defence functions. The asset configuration information as well as the vulnerability information are described using OVAL[2] and CVE[3] respectively. The automated, AI aspects of the self-healing component includes a collection of security policies, rules and mitigation measures.

3. Evaluation framework

When handling cyber security incidents, a Security Operations Centre (SOC) operator can be faced with the following challenges and problems [6, 7, 8, 9, 10]:

- Complexity: Cyber incident response systems can be complex and require a high level of technical expertise. This can be challenging for human operators who may not have the required knowledge and skills to navigate the system effectively.
- Time pressure: Incident response requires quick action to prevent further damage, which can create time pressure for human operators. This can lead to mistakes or oversights, especially if they are not trained to work effectively under such pressure.
- Alert fatigue: Cyber incident response systems generate numerous alerts, which can be overwhelming for human operators. This can lead to alert fatigue, where operators become desensitised to alerts and may miss critical information.

[1] https://opensearch.org/

[2] https://oval.mitre.org/

[3] https://cve.mitre.org/

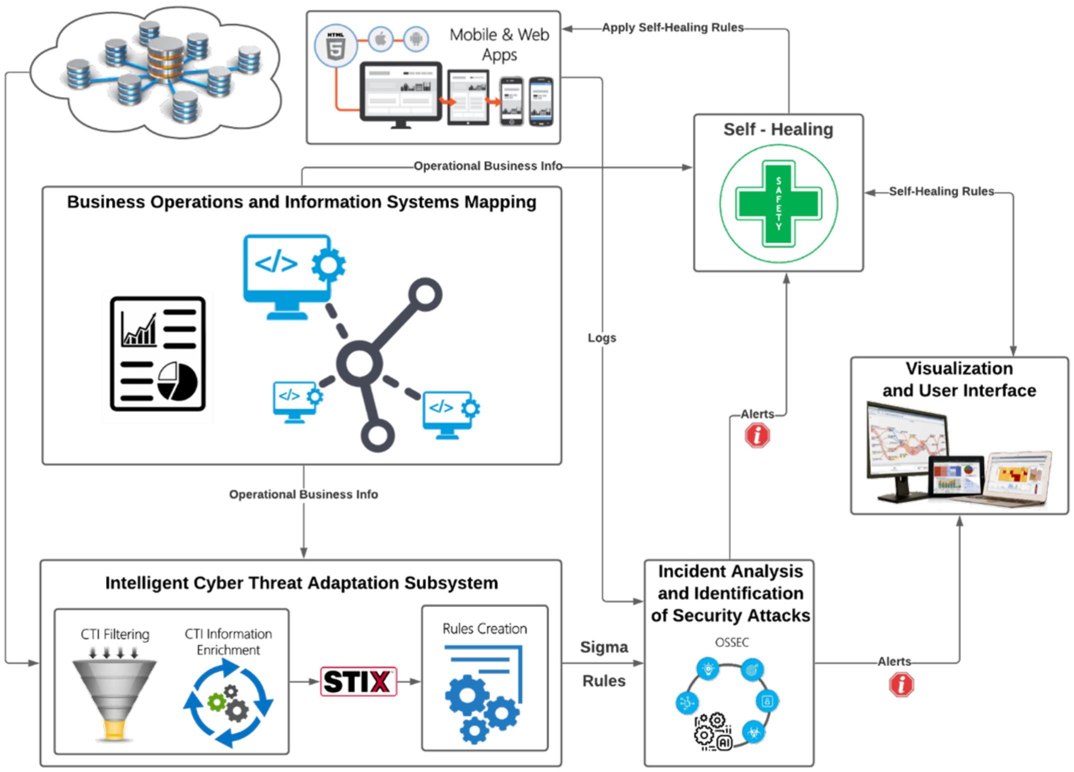
Mobile & Web Apps
Apply Self-Healing Rules
Self - Healing
SAFETY
Operational Business Info
Business Operations and Information Systems Mapping
Self-Healing Rules
Logs
Visualization and User Interface
Alerts
Operational Business Info
Intelligent Cyber Threat Adaptation Subsystem
CTI Filtering
CTI Information Enrichment
STIX
Rules Creation
Sigma
Rules
Incident Analysis and Identification of Security Attacks
OSSEC
Alerts

Figure 1: The proposed CTI sharing and response architecture.

- Lack of integration: Cyber incident response systems often work in silos, which can make it difficult for human operators to integrate information from different systems. This can result in incomplete or inaccurate incident response.
- Lack of automation: Some incident response processes can be automated, but many require human intervention (e.g. scanning network traffic for malicious activity, analysing logs for suspicious behaviour, quarantining infected systems). This can be challenging for human operators who may need to handle multiple incidents at once, leading to delays or errors.
- Lack of training: Human operators may not be adequately trained on the cyber incident response system, leading to ineffective use of the system and potentially leaving the organisation vulnerable to attacks.

The proposed integrated incident detection and response system can be evaluated on two fronts:

- **User Experience (UX) based.** This refers to the approaches dealing with the assessment of users' needs, behaviours, attitudes, and preferences when interacting with a product,

service, or a system. When considering SOC analysts and their additional mental and cognitive load when dealing with security incidents, assessment frameworks such as NASA's Task Load Index (NASA TLX) [11] can be employed to evaluate the perceived workload of an individual or a team during a task.
NASA TLX is a subjective measure and consists of 6 sub-scales [11] which can be used to evaluate a SOC participant when engaging with an incident detection and response system as follows:

  - Mental demand. This refers to the mental effort required to identify and analyse threats. For example, analysing network traffic patterns to detect anomalous behaviour, or researching new attack vectors and techniques used by threat actors.
  - Physical demand. The physical effort required to implement and maintain security controls. For example, deploying and configuring firewalls, updating and deploying rules, and performing regular maintenance and updates.
  - Temporal demand. The time pressure or urgency involved in detecting and responding to cyber attacks. For example, rapidly detecting and responding to malware infections – such as ransomware – or data breaches to minimise the impact on the organisation.
  - Performance. The perceived quality of security performance. For example, measuring the effectiveness of security controls in preventing or mitigating cyber attacks, or assessing the accuracy and reliability of cyber threat intelligence data feeds.
  - Effort. The overall level of effort required to implement and maintain effective cybersecurity measures. For example, investing in robust security tools and technologies, developing and implementing security policies and procedures, and providing ongoing training and education for security personnel.
  - Frustration. The level of frustration or stress experienced by security personnel during the detection and response phases. For example, dealing with false positives or false negatives from security tools, or managing the workload and stress of responding to multiple security incidents at the same time.

- **Human in the Loop driven.** When deploying any system with a substantial machine learning or AI component, the Human in the Loop (HITL) aspects should be considered. In the context of SOC analyst or operator, we argue that HITL can vary depending on the degree of uncertainty surrounding an attack. By taking uncertainty into consideration, the CTI information will serve a better purpose and be more actionable. In this work we consider the Cynefin framework [12] that can be used to guide decision-making and problem-solving during cybersecurity incidents. Cynefin has five so-called dimensions or contextual definitions that can help SOCs to develop a more nuanced and flexible approach to cybersecurity, taking into account the specific characteristics of different types of threats and attacks. The dimensions and an example of how they could be applied to incident handling are as follows:

  - **Simple (known knowns)**. In the simple dimension, problems are well-defined, and there is a clear cause-and-effect relationship between the problem and the solution. In the context of cyber attacks, the simple domain could be applied to routine security tasks such as patch management, security configuration, and access control. Indicators of Compromise (IoCs) are unambiguous and can attribute the threat.

- **Complicated (known unknowns)**. In the complicated dimension, problems reach a state where there may be multiple potential solutions that require expert knowledge and analysis. In the context of cyber attacks, the complicated domain could be applied to tasks such as incident response, malware analysis, and vulnerability assessments. IoCs are somewhat unambiguous and can could attribute the threat with a bit of effort.
- **Complex (unknown unknowns)**. In the complex dimension, problems are unpredictable and emergent, and there may be no clear cause-and-effect relationship between the problem and the solution. In the context of cyber attacks, the complex domain could be applied to threat hunting, threat intelligence, and adaptive security measures. IoCs are ambiguous and not as trustworthy.
- **Chaotic (unknowables)**. In the chaotic dimension, problems are unpredictable and rapidly changing, and immediate action is required to stabilise the situation. In the context of cyber attacks and the kill chain, the chaotic domain could be applied to the initial response to a major cyber incident, where there is a need for rapid triage, containment, and recovery. IoCs cannot be defined, or if they do so, they have almost no value as they will be too generic or not trustworthy.

Having established a structured framework of *sense making* when analysing and responding to cyber security incidents, the SOC analyst will be in a position to navigate through different response options while managing their expectations of the participation of the AI/self-healing layer as well as their intervention. In Figure 2 an example mapping of the mix between automated response and human participation across the cyber kill chain is presented. Assuming that advanced threats deliver campaigns comprised of a number of carefully sequenced attacks, the risk and impact of the attack in principle increases as the adversary progresses within the kill chain. For each progression it is assumed that at least one asset has been compromised, which in turn suggests that the security controls were not effective. As such, the need for human intervention and participation is expected to increase.

Moreover, considering the level of uncertainty (and complexity) as expressed by the Cynefin framework, we expect that the higher the degree of uncertainty, the more demand of human intervention, at an earlier stage of the kill chain. This is required as in high uncertainty situations it may not even be possible to distinguish on which phase of the kill chain a (detected) attack will be at.

4. Concluding remarks and ongoing work

In this paper we sketched an approach for assessing and studying the challenges that arise from introducing AI-facilitated operations – that is, self-healing – in the cyber incident handling lifecycle. This approach fuses subjective measurable features and dimensions that are of significance to the interplay of the AI and human operator interaction. Using the proposed architecture developed for the Cyber-pi project, the future research will develop and deploy

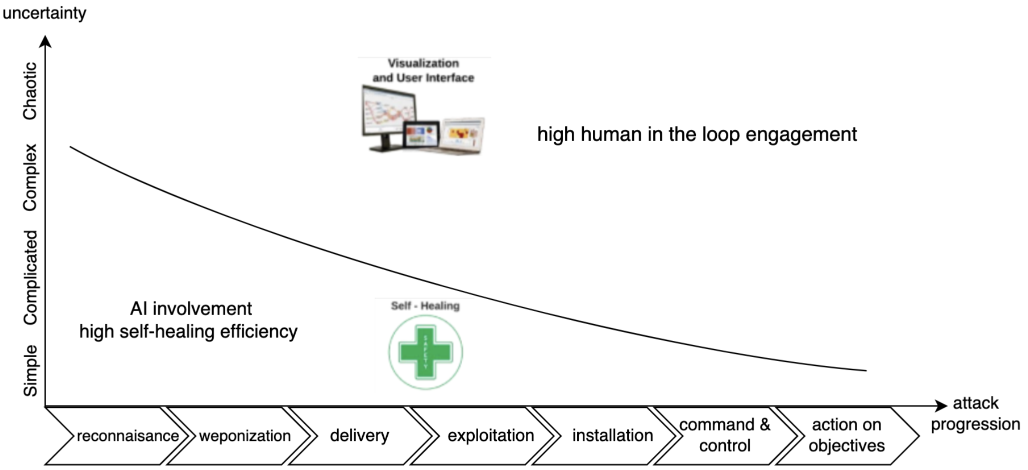

Figure 2: AI to human involvement.

cyberthreat scenarios and use cases that will be deployed on a cyber range. This will enable the empirical assessment and evaluation of the approach.

Acknowledgments

Co-financed by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH - CREATE - INNOVATE (project code: T2EDK-01469).

References

[1] M. Spremić, A. Šimunic, Cyber security challenges in digital economy, in: Proceedings of the World Congress on Engineering, volume 1, International Association of Engineers Hong Kong, China, 2018, pp. 341–346.

[2] B. Al Sabbagh, Cybersecurity incident response: a socio-technical approach, Ph.D. thesis, Department of Computer and Systems Sciences, Stockholm University, 2019.

[3] K. Kandasamy, S. Srinivas, K. Achuthan, V. P. Rangan, Iot cyber risk: A holistic analysis of cyber risk assessment frameworks, risk vectors, and risk ranking process, EURASIP Journal on Information Security 2020 (2020) 1–18.

[4] A. Malik, Microsoft says Bing can be provoked to respond outside of its 'designed tone', https://techcrunch.com/2023/02/16/microsoft-bing-provoked-respond-outside-of-designed-tone/, 2023. [Online; accessed 22-February-2023].

[5] S. Halangoda, OpenAI ChatGPT for Cyber Security, https://infosecwriteups.com/openai-chatgpt-for-cyber-security-4bc602069f9c, 2022. [Online; accessed 22-February-2023].

[6] E. Agyepong, Y. Cherdantseva, P. Reinecke, P. Burnap, Challenges and performance metrics for security operations center analysts: a systematic review, Journal of Cyber Security Technology 4 (2020) 125–152.

[7] C. Zhong, T. Lin, P. Liu, J. Yen, K. Chen, A cyber security data triage operation retrieval system, Computers & Security 76 (2018) 12–31.

[8] P. Lif, T. Sommestad, Human factors related to the performance of intrusion detection operators., in: HAISA, 2015, pp. 265–275.

[9] L. Aijaz, B. Aslam, U. Khalid, Security operations center—a need for an academic environment, in: 2015 World Symposium on Computer Networks and Information Security (WSCNIS), IEEE, 2015, pp. 1–7.

[10] S. C. Sundaramurthy, M. Wesch, X. Ou, J. McHugh, S. R. Rajagopalan, A. G. Bardas, Humans are dynamic-our tools should be too, IEEE Internet Computing 21 (2017) 40–46.

[11] S. G. Hart, NASA task load index (TLX), Technical Report 20000021488, Human Performance Research Group, NASA Ames Research Center, Moffett Field, California, 1986.

[12] D. J. Snowden, M. E. Boone, A leader's framework for decision making, Harvard business review 85 (2007) 68.