Sharing and Automation for
Privacy Preserving Attack Neutralization

(H2020 833418)

D4.6 Algorithm to automate recommended response and recovery actions without human operators, first version (M15)

Published by the SAPPAN Consortium

Dissemination Level: Public

H2020-SU-ICT-2018-2020 – Cybersecurity

Document control page

Document file: **D4.6 Algorithm to automate recommended response and recovery actions without human operators, first version (M15)**
Document version: 1.0
Document owner: Avikarsha Mandal (FIT)

Work package: WP4
Task: T4.4
Deliverable type: Other
Delivery month: M15
Document status: ☒ approved by the document owner for internal review
☒ approved for submission to the EC

Document History:

Version	Author(s)	Date	Summary of changes made
0.1	Avikarsha Mandal (FIT), Lasse Nitz (FIT)	2020-07-24	First Draft
0.2	Avikarsha Mandal (FIT)	2020-07-27	Draft Ready for review
1.0	Avikarsha Mandal (FIT), Lasse Nitz (FIT)	2020-07-31	Reviews collected and incorporated. The deliverable is ready for submission.

Internal review history:

Reviewed by	Date	Summary of comments
Franziska Becker (USTUTT)	2020-07-27	Spelling, grammar and sentence structure
Mischa Obrecht (DL)	2020-07-29	Spelling, structure, content
Martin Zadnik (DL)	2020-07-31	Technical

Executive Summary

This deliverable is the first version of SAPPAN approaches for Automatic Response and Recovery steps without the involvement of human operators. In this task, approaches for automating response and recovery steps without the involvement of human analysts will be developed and evaluated. The intrusion detection tools usually generate a large amount of false-positive alarms that are handled manually by human analysts to take appropriate response action. As a limited number of human operators are available for the analysis of too many false alarms, it is possible they could miss some true alerts due to fatigue/inexperience, which can have hefty consequences. The goal of this task is to develop an approach that takes away some of the burdens of human operators to rapidly react to potential attacks in real-time, while also decreasing the damage which the wrong response at the wrong time can cause. In this SAPPAN approach, we contribute towards automating some types of response actions depending on the confidence level associated with the detection, the importance of the asset involved, and the assessment of the severity of an incident. Additionally, we contribute towards developing approaches to measure the risk of automatically performing an action in case of a false positive versus the risk of missing the potentially very narrow window of time to mitigate or contain an attack.

This deliverable D4.6 is part of Task T4.4, where we explore approaches for automating response actions for specific incidents/attacks without human analysts. In this preliminary version, we provide the necessary background and related works in the research direction of automating response and recovery steps for cybersecurity incidents. As part of this deliverable D4.6, we propose a framework that can capture the approaches and algorithms for automating some types of responses actions depending on several factors. We select two showcases from WP3 that have moderate risks for response actions and discuss the steps of automation in line with our proposed framework. Further development and concrete evaluation will be part of the final deliverable D4.7.

Contents

1 Introduction

Cybersecurity incidents and cyberattacks can cause serious damage to any type of organization. To effectively mitigate such attacks, response and recovery steps play an important role in incident management. There exists a large number of detection modules on the cybersecurity market that generate large amounts of events and warnings. In common practice, when attacks are detected, the response actions are taken manually by experienced human analysts. However, a serious challenge is during post-detection where cyber threats must be efficiently and effectively handled with only a small number of expert security analysts available to interpret massive amounts of data. The selection of a response or the recommendation of a response heavily depends on the type of the detected attack, confidence-level of the detected attack, involved assets, assessment of the severity of an incident, etc.

There are different types of response actions for different types of incidents such as host-based, network-based, hybrid and distributed [1] [10]. Some of the response actions that can be triggered on a host typically include the termination of a process, taking backup of the system (snapshots such as processes running their capabilities, network connections, registry, etc.), process isolation from the network, hosts isolation from the network or from the internet, shutting down infected hosts, sending out messages to end-users, deleting/locking files on a host. Response actions that can be performed on the network side include: blocking traffic to/from particular IP addresses, port block, TCP connection reset, access control lists chaining, reconfiguring routers and firewalls, etc. Further examples for responses include infected node isolation, IP address relocation to a not-infected server, locking user accounts, disconnecting network, creating backups, or triggering more in-depth monitoring.

Whereas automated response recommendations to the human operators help them to select appropriate response action for specific attacks (in progress within Task T4.3), fully automated response and recovery actions will reduce the overhead of human operators significantly for less critical and well-known incidents. In recent years, Intrusion Detection Systems (IDS) with semi-automated to automated responses became an essential research area from an Industrial perspective [1].

In the next sections, we provide the necessary background and related works in the research direction of automating response and recovery steps for cybersecurity incidents. Furthermore, we aim to propose a general framework for SAPPAN that can capture the approaches and algorithms for automating some types of responses actions depending on several factors. We identify two showcases from WP3 (Phishing and Domain Generation Algorithm (DGA)) that have moderate response risks and discuss the steps of automation in line with our proposed framework.

2 Background for Automating Response and Recovery

In the course of SAPPAN, an in-depth literature review for automated response and recovery actions has been performed by Burian in [2] and some parts of this section are based on Burian's work.

2.1 Overview of Incident Management

2.1.1 Incident Response and Recovery Process

Figure 1: Incident Response and Recovery Process

The incident management process and its workflow vastly vary from organization to organization. Usually, Computer Security Incident Response Teams (CSIRTs) or Computer Emergency Response Teams (CERTs) are responsible for the handling of cybersecurity incidents. A typical incident response and recovery process is shown in Figure 1 from different guidelines [3][12][13] and mainly consists of 4 phases:

- **Preparation:** This phase includes the preparation for handling incidents (usually documented in the organization's cyber exercise playbooks [14]) and organizational measures to prevent specific incidents from happening. For the preparation for handling incidents, the security operations center performs different activities such as preparation of facilities, communications, hardware, software incident reporting mechanisms, issues tracking system, network security monitoring tools, digital forensics software, and mitigation software. For the prevention of incidents, it contains risk assessment processes, host and network security mechanisms, malware protection, and user awareness and training, etc.

- **Detection:** In this phase, the actual detection of a cyber threat takes place. The sources for detections include, but are not limited to, Intrusion Detection/Prevention System (IDS/IPS), Security Information and Event Management (SIEM) systems, antivirus/antispam software for malware detection, network logs, application logs, public information (e.g., National Vulnerability Database (NVD)),

etc. In general, detection systems use signature-based or statistical methods to identify malicious activities.

- **Assessment:** The assessment phase of the response and recovery process aims to find the root cause of the incident with data analysis, incident assessment, incident prioritization, and documentation. First, data analysis of events is performed from available data sources identified in previous steps from the security perspective. It is still challenging for many organizations to determine whether an incident has occurred or not, as there are a large number of false positives. Experienced security analysts consider several aspects to mark a set of events as an incident and prioritize accordingly, and document assessment results for future events.

- **Handling:** The handling phase contains the core *response and recovery* steps. The involved steps during the handling phase usually are:

    - **Containment Strategy Definition:** The aim of the containment strategy is to isolate or mitigate the threat before it increases damage to the infrastructure. The strategies for this step highly vary depending on the incident type. For example, a strategy for containing a DDoS attack (such as load balancing or making use of a scrubbing center) is different from the containment strategy of malware infection (e.g., isolation of an infected computer).
    - **Evidence Gathering:** Here, evidence of an incident is collected for handling (e.g., technical, victim testimony). This is also required for legal proceedings.
    - **Identification of Attacking Hosts:** Sometimes, system owners or operators want to identify the attacking host. To do this, common activities are validating the attacker's IP address, researching in search engines, or monitoring probable communication channels used by the attacker.
    - **Eradication and Recovery:** Eradication is required in some cases to eliminate the consequences of the incident (e.g., deleting malware or identifying and mitigating all vulnerabilities that were exploited). Furthermore, it is important to identify all affected hosts which can be remediated. In the recovery step, systems are restored to normal operation. If it is possible, vulnerabilities are remediated to prevent similar incidents. Recovery may include several operations such as taking clean backups, freshly rebuilding the system, replacement of compromised files, changing passwords, etc.
    - **Lessons learned:** All newly discovered knowledge captured in the previous steps should be documented and collected for the next preparation phase for handling similar incidents in the future.

2.1.2 Standards

To understand the incident management process and best practices, we outline some standards and guidelines for incident-related activity as follows:

- **ISO 20071** [6] is a series of standards developed by ISO (International Organization for Standardization) and IEC (International Electrotechnical Commission) for incident management. The incident management process in this standard is known as "Management of information security incidents and improvements". The process follows several phases respectively: Responsibilities and procedures, Reporting information security (IS) events, Reporting IS weaknesses, Assessment of and decision on IS events, Response to IS incidents, Learning from IS incidents, Collection of evidence. This standard is widely used and maintained by more than 27000 companies in 2015.

- **Information Technology Infrastructure Library (ITIL)** [15] is a collection of usual processes and workflows inside an organization aggregated under the topic of “IT service processes”.
  Regarding incident handling, the ITIL-R Respond Workflow ([7], p. 221)) compared to ISO 27001 has a clearer defined process of incident response.

- **NIST Special Publication 800-61** [13] is a set of incident handling guidelines published by the National Institute of Standards and Technology (NIST). They suggested three main activities in the incident handling phase: i) “Lessons Learned” to be help on a regular basis; ii) “Using collected Incident Data” such as the number of incidents handled, time per incident. objective assessment, subjective assessment; and iii) Audits based on previous handling of incidents.

- **SANS Incident Handler's Handbook** [16] is a collection of practice-oriented checklists and templates which has its own definition of incident management. The incident handling has six phases as preparation, identification, containment, eradication, recovery, lessons learned.

2.1.3 Incident Information Sources and Platforms

For incident response and recovery, it is important that organizations have a good and reliable source of information about currently known threats and standard response actions. We provide some information sources for incidents which might be useful in the context of this task:

- **MITRE ATT&CK-based Analysis** [17] consists of a data source and an analytic process, primarily for the detection of advanced threats (e.g., attacks or hacks by any individual or group). The project has two main objectives: i) emulate attacks as an attacker would play them out (known as adversary emulation playbook technique), and ii) analyzing an organization’s network and discovering security coverage and gaps within the network. Note that ATT&CK-based framework might not be suitable for approaches to mitigate attacks but rather a tool for the preparation phase.

- **MITRE Common Vulnerabilities and Exposures (CVE)** [18] is a list of entries, comprising descriptions and references for cybersecurity vulnerabilities that are publicly known. The CVEs are not incidents but they list a wide-range existing vulnerability that represents a potential threat.

- **MISP (also known as Malware Information Sharing Platform)** [23] is an open-source threat intelligence platform that can be used for collecting, storing, sharing cybersecurity threat information. Specifically, the platform provides a database of incident indicators and indicators of compromise (IoC), a correlation engine, and an event graph functionality to visualize the relationship between different attributes and objects. Furthermore, MISP offers an API that can be used for automation and data sharing.

- There are few computer security threat exchange platforms such as *circl.lu, AusCERT, US-CERT, mycert* [19-22] to share knowledge about cybersecurity incidents. These platforms issue notifications/bulletins)regularly, some for their members, and some for the public.

2.2 Intrusion Response System (IRS)

The Intrusion Response System (IRS) and its components are particularly designed to identify and mitigate potential incidents efficiently as a security countermeasure [10]. There exist detailed taxonomies for IRS in [1][8][10]. Normally, the Intrusion Detection System (IDS) only sends an alarm to the operator if any malicious activity is detected in the network. In contrast, the Intrusion Prevention System (IPS) mainly deals with preventive measures before an incident takes place. Lastly, the Intrusion Response System (IRS) offers response capabilities in order to mitigate attacks. The IRS can be classified based on the selection of the response method, level of automation, cost of response, etc. In a broader sense, IRS based on triggered responses can be divided into passive and active. Passive IRS include notification-based/manual response and recovery actions where victims are notified with alarms/email in response to some detected incidents by some human operators. In this type, there is a delay between the detection time and the time when an alert is sent. The active IRS includes fast and automated response actions without the involvement of human operators. Automated IRS consists of three types [10] such as i) Adaptive-based system, ii) Expert-based system, and iii) Association based system. In adaptive-based automated IRS, there is a feedback loop to evaluate previous responses. In Expert-based automated IRS, the decisions regarding response actions are based on one or more metrics. In Association-based IRS, there is a decision table where each response action is linked with a specific attack.

2.3 Related Works in Response Automation

From the perspective of current industry offering and academic literature, an in-detailed review is provided in the SAPPAN deliverable D4.4.

3 Design of SAPPAN Framework for Automating Response and Recovery Actions without Human Operators

In this section, we design a conceptual framework to capture SAPPAN approaches and algorithms for automating some response actions without human operators. To do this, we provide some high-level requirements, our approach for automated responses without human analysts in comparison with any typical manual response and response recommendation, and finally involved steps we plan to follow.

3.1 Requirements

The first step in developing an algorithm for automated response actions is to collect general requirements. we identify several factors for the feasibility of automating response actions:

- **Incident type:** There exists a wide variety of intrusions/attacks such as phishing, DDoS, Malware, and many more. Depending on the incident type, the response actions are vastly different. Whereas some response actions are relatively simple and possible candidates for automation, others can be highly complex and might be too risky/infeasible to automate without the involvement of human analysts. Hence, it is important that the automation system can identify which response actions to carry out automatically, depending on the incident type.
- **Precision of the detection:** If a specific incident can be detected with higher precision, the response steps for that incident might be a good candidate for full automation. If true alerts can be recognized automatically with very high certainty, automatic response steps might be done without human operators.
- **Importance of the asset involved:** Asset risk management is an important aspect to consider, such that damage that can incur from the incident or the damage of carrying out a wrong response action can be minimized. The response actions for incidents affecting low-cost assets might be carried out automatically.
- **The assessment of the severity of an incident:** Assessment of the severity of the incident is a critical factor. For example, IBM cloud event management systems [11] proposes a prioritization of 5 severities, where priority level 5 represents the lowest severity level (e.g., the possibility of data expose which are publicly available/not critical), while priority level 1 has the highest severity level is (e.g., successful attacks on company-internal infrastructure). Less severe incidents might be preferable for an automated response action.
- **Risk of response action:** The time frame involved to mitigate an attack is an important factor. Some attacks are time-critical and response actions for mitigation should be carried out immediately. Furthermore, automatically performing some response actions for false-positive alerts can have severe consequences. In the SANS Institute study [9] conducted among their organizations, Bromiley found that 42% of incidents were not handled within the first 24 hours after detection and mitigating another 22.5% of the incidents can take up to 5 to 24 days. Therefore, the expected time to handle an incident is a deciding factor for the automation of response actions. Hence, the framework needs a mechanism for quantifying the risk of automatically performing an action in case of a false positive, versus the risk of missing the potentially very narrow window of time to mitigate or contain an attack.

- **Cost of response action:** The cost of the response action might be an important factor to design automated response algorithms (e.g., cost-sensitive IRS [10]). As mentioned in [1][10], the cost of response action must be lower than the cost of damage caused by the incident. A response action can get triggered when the damage cost is greater than the response cost.

3.2 Transition from Manual to Automated Response and Recovery

To capture our approach for automation, we reuse some of the SAPPAN Architecture components proposed in WP2 (Deliverable D4.2).

3.2.1 Framework Components

To capture our approach for automation, we reuse some of the SAPPAN Architecture components proposed in WP2 (Deliverable D4.2).

Component Name	Description
Human Analyst	Human analyst takes the decision for specific response actions for detected incidents. His judgment is used to label incidents in the training data, to update the ML model, and to trigger response actions.
Unseen Incidents	New incoming incidents being checked for attack.
Past Incidents/Training Data	It contains datasets/past incidents/features that are used for machine learning to detect new incidents. The training data contains datasets/features that are used for training ML models, having the human analyst involved in the process of adding new samples to the training data allows for labeling these samples.
SIEM	A security information and event management (SIEM) system that aggregates and displays alerts for an organization. The SIEM database is used as the knowledge-base to store the results of the detection system or Machine Learning Engine and provide them to the human analyst. The database may contain detection alerts, as well as related primary data or alert severity. The database is also able to correlate individual events into complex events and provide aggregated information about detected attacks or anomalies.
Detection System/ML Engine	Detection System/ML Engine is a technical component that executes machine learning tasks for detecting incoming incidents. The ML engine is able to train models based on many types of machine learning architectures using different libraries (e.g. TensorFlow or PyTorch). It can also update the ML model based on feedback from human analyst/automated response and recovery system.
Decision Support System	Decision support system supports the human analyst for taking decision regarding response actions.

Automated Response and Recovery System	Automated response and recovery system upgrades decision support system such that low-risk response actions can be taken without human analysts.

3.2.2 Manual Response and Recovery:

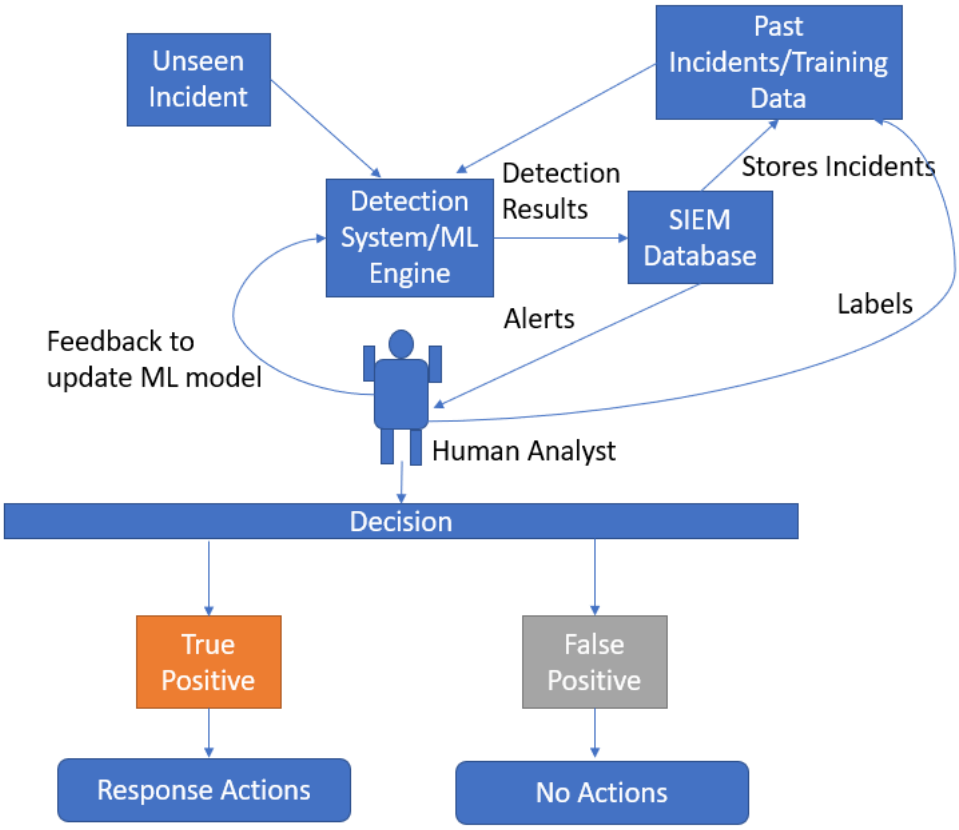

Figure 2: Process of Manual Response and Recovery

To design an automated response and recovery framework, the first step is to understand some standard approach of manual response and recovery process by using SAPPAN components. The majority of the intrusion response systems handle incidents manually with some human analysts. When new incidents/attacks are detected by the Intrusion Detection System (e.g., ML-based detection engine), the human analyst in the operation center receives a detection notification in the form of alerts. Then the human analyst manually separates false positives to identify real incidents, considering several factors and using prior experience. These decisions are then used to label the respective entries in the training data. For false-positive cases, the analyst declares the result as false positive to the detection system and no actionable response will be taken. For the true positives, the analyst triggers the response actions (notification or manual [1][10]). For low-risk incidents and response actions, the operator sends an alert/report to the victim via email or some other form of notification (*notification based response*). For *manual response system*, operators have some pre-defined

rules of response actions (e.g., following playbooks) and get triggered with true positive detection. One of the weaknesses of notification and manual response systems is the time duration between detection and response action, which gives the attacker more time to do further damage with high-speed active attacks.

3.2.3 Semi-Automated Response and Recovery:

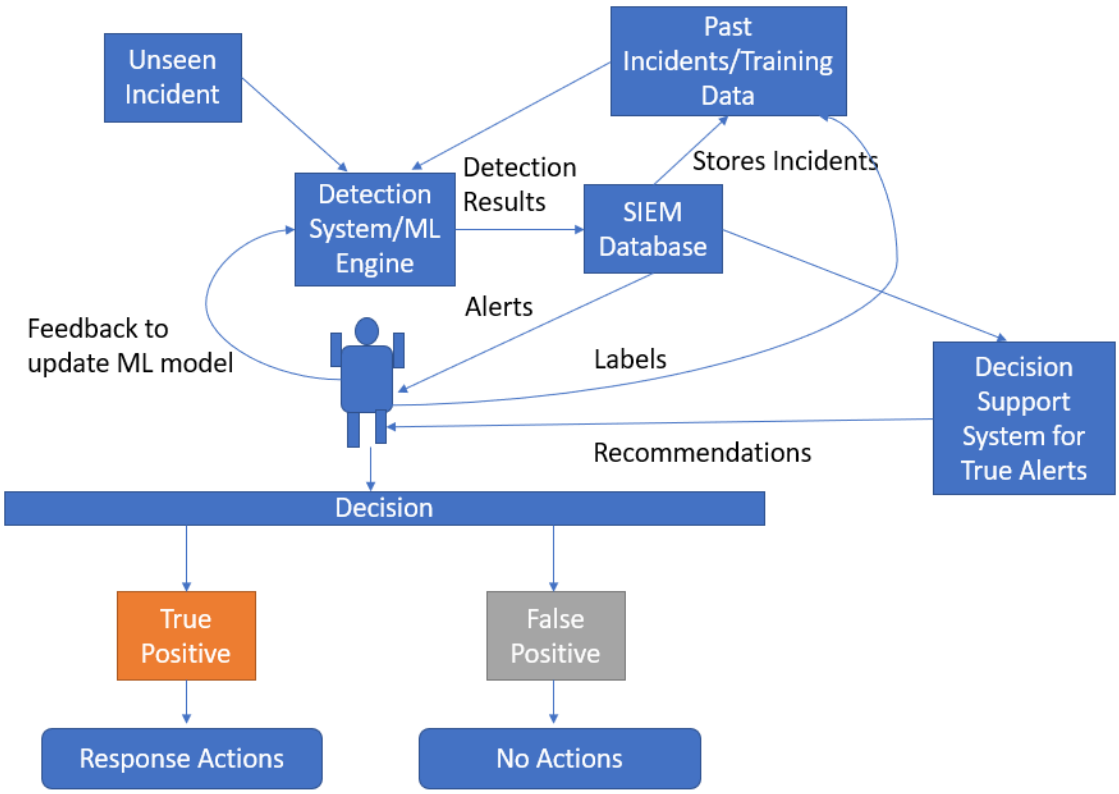

Figure 3: Process of Semi-Automated Response and Recovery

The next step towards a fully automated response and recovery process is designing a semi-automated system which offers response recommendations and reduces false positives. As false positives alerts can outnumber true alerts by a factor of 10 or even more, automated approaches of response recommendation to filter out false-positive will help the human analyst significantly. Architecture-wise, an additional decision support system is required to recommend true alerts to the human analyst. Depending on the type and severity of a true incident, an alert can suggest to human operators to take appropriate response actions, or in certain cases, it might be possible to execute the response automatically. In the context of SAPPAN, task T4.3 is currently investigating different machine learning-based approaches such as incident similarity model, host aggregation similarity model, and false alert recognition to build mechanisms to automatically identify false positives.

3.2.4 Conceptual Framework for Automated Response and Recovery in SAPPAN:

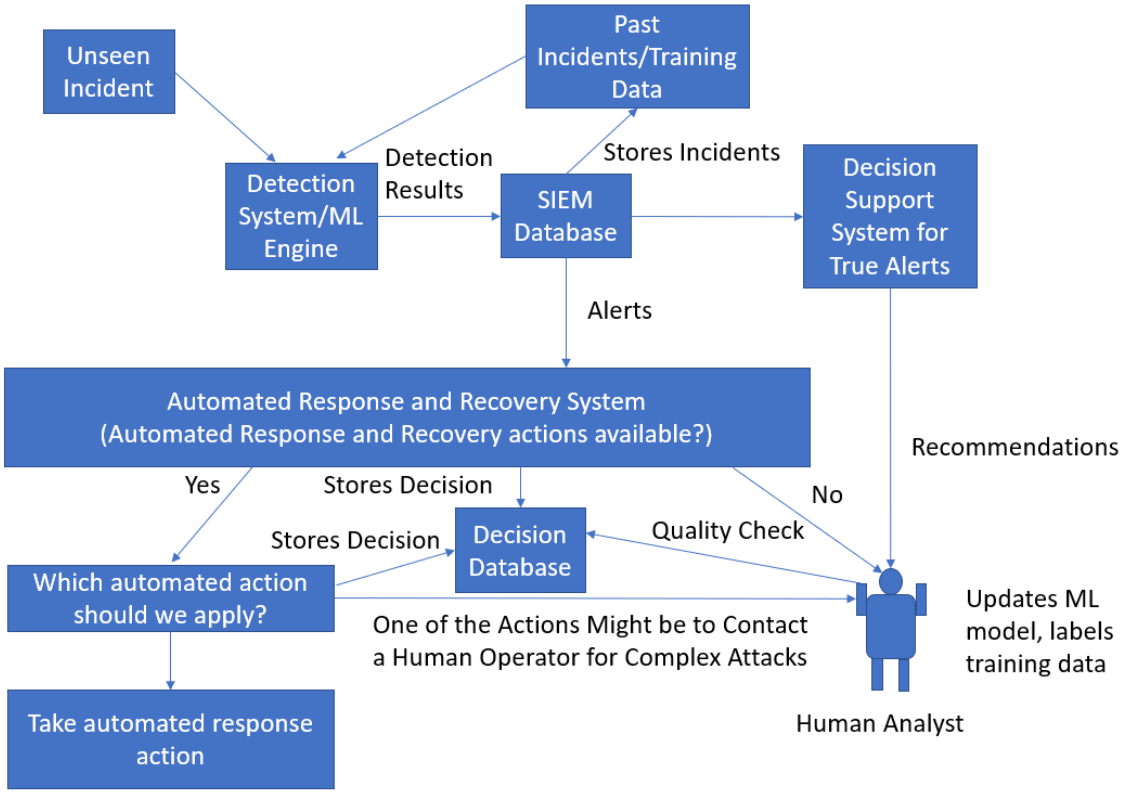

Figure 4: Process of Automated Response and Recovery in SAPPAN

The SAPPAN approach for automating response and recovery steps is shown in Figure 4. Our main focus is to explore automation algorithms for response actions without the involvement of a human operator. In our general architecture, we kept a human analyst in the loop in case of particular response actions that are not suitable for automation. We assume our automated response and recovery system can identify false incident alerts with very high confidence. The system checks if there are any response actions suitable for automation depending on different factors such as incident type, risk of response, etc. In the case of negative results, the alert is simply forwarded to another human analyst and the human analyst makes a decision with the help of the recommendations provided by a decision support system. For the positive cases, the main research question is: *which response actions can be automated efficiently and how can the system trigger some response action without the involvement of human operators?* In certain cases such as complex attacks, our framework can allow contacting human analysts as one of the response actions. In our architecture, we include a decision database that stores decisions taken by the automated system for monitoring the automated system, deriving new training data, and possibly additional investigation. Note that the current framework components are subject to change for the final version. In the later section, we discuss one potential approach to design some fully automate response actions based on *confidence score*, *certainty metric, and response threshold.*

3.3 SAPPAN steps for automating response and recovery action without human operators

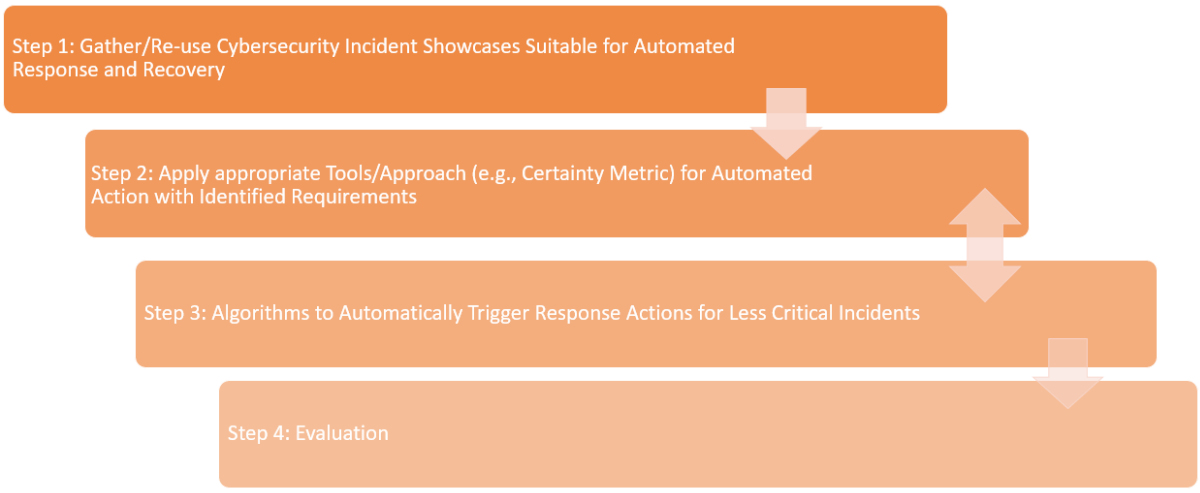

Figure 5: Steps for Task T4.4

As shown in Figure 5, we plan to follow involved steps to successfully perform the task T4.4. In the first step, we select some showcases which we believe have some low/moderate risk of response actions and good candidates for automated response actions. At this stage, two showcases from WP3 (phishing and DGA) have been selected and more showcases might be considered for future evaluation. In step 2, we want to investigate suitable approaches/tools for automation of response actions with identified requirements. For example, we might evaluate several certainty metrics for different attack types to quantify some confidence scores for attack detection. Moreover, different techniques from the recommender systems domain [4][9] or Case-based reasoning [2] might also be suitable for this step, depending on the incident type. In the next step, we propose mechanisms that will trigger automated response actions for suitable incidents. For example, we can define thresholds for different confidence scores to trigger different response actions. Finally, some quantitative evaluation of the implemented algorithm must be performed.

4 Risk Assessment to Measure the Severity of an Incident and Mitigation

Risk assessment while designing an automated response approach is an important aspect. The response actions are highly dependent on the criticality of the asset involved (network, infrastructure, information) which is under an attack. Due to a possible incident, the cost of damage involving different assets is not equivalent and the respective response action might not be cost-effective for the organization. In our case, two types of risk assessments might be required [10]:

- **Assessment of Incident:** Risk assessment approach to evaluate an incident can be Attack graph-based, Service-dependency graph-based, Non-graph based.
- **Assessment of Response Action**: To design our system for automated response, different aspects such as success rate of previously deployed response, potential damage while applying new response, response cost, etc.

must be analyzed from a risk assessment perspective. Moreover, there are two types of risk assessment mechanisms that are commonly applied to response execution i.e. Burst and Retroactive.

Additionally, as an alternative to risk assessment approach (or to get supplementary information), we might conduct interviews with domain experts to find suitable candidate showcases for response automation without human operators.

5 Confidence Score for Incident Detection and Thresholds for Response Actions

As part of this chapter, we list considerations which may serve as a basis for the definition of a general framework. The main focus lies on two aspects of central importance: A confidence score that quantifies how certain the system is that a certain type of attack occurred, and the definition of thresholds to trigger associated response and recovery actions. It is of key importance that these two aspects are well attuned, since the system should try to minimize negative impact of response and recovery actions in case of a false-positive detection. The definition of the confidence score and the response and recovery thresholds are defined per type of attack. This allows to consider that different sub-systems are used for different kinds of attacks, and that the quality of the detection of these sub-systems may vary. As a consequence, it might also be that different sub-systems detect different threats with comparable confidence. The problems and possibilities introduced by this are also discussed as part of this chapter. Figure 6 shows an abstract overview of how the confidence score and the response and recovery thresholds interact.

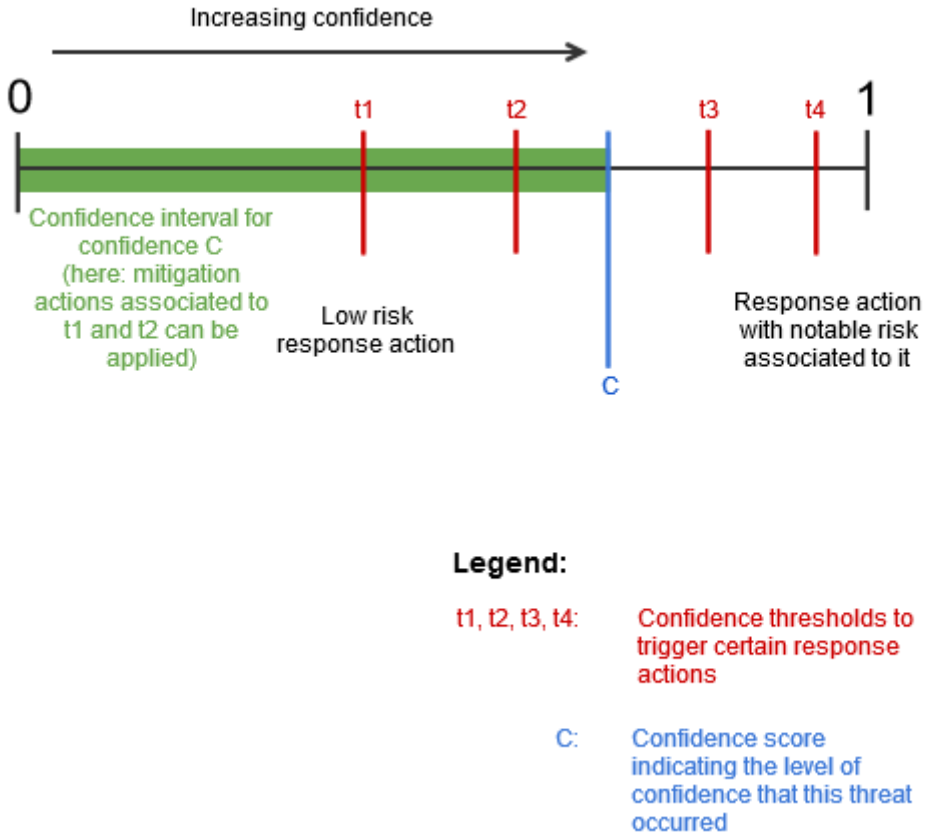

Figure 6: Diagram showing the different components that are associated with the confidence score for one type of attack. The confidence score C quantifies how certain the system is that this kind of attack occurred. The four different thresholds t1, t2, t3, and t4 have response and recovery actions associated with them. A specific response and recovery action can be carried out if the confidence is higher than the threshold of this response and recovery action. In the diagram, this is visualized by the green confidence interval. The higher the confidence threshold for a certain response and recover action is, the higher the damage is in case of a false-positive classification of the threat.

5.1 Confidence Score

The confidence score serves as a quantification of how confident the system is that a certain type of attack has occurred. For clarification, let the term *confidence score* refer to the level of confidence that a given type of attack occurred, and let the term *certainty metric* refer to an actual function that could be used to compute the confidence score. Consequently, the confidence score is computed via a specific certainty metric, and different certainty metrics could be used to compute the confidence score of different attack types.

It is preferable that all certainty metrics have the same domain (such as [0,1]), since this would allow to compare the confidence levels of different kinds of attacks and to change certainty metrics without applying changes to other components. Thus, already defined thresholds do not need to be changed necessarily. Especially for the purpose of evaluating the suitability of different certainty metrics, this property allows for direct comparison of results.

Because the confidence score is utilized to take specific response and recovery actions, the definition of suitable certainty metrics is one of the crucial aspects of task T4.4. In the following, some factors that should be considered in the definition of suitable privacy metrics are presented.

- The false-positive rate of the detection system
  - A high false-positive rate implies low trust in the quality of the detection and, thus, low confidence that an attack occurred in case of a detection
- It might be necessary to consider a sequence of detections to improve confidence
- The value range of the certainty metrics needs to match the thresholds for mitigation actions

5.2 Response Action Thresholds

Defining thresholds for response and recovery actions is one of the crucial aspects of automating respective actions based on confidence. One key observation is that not all response and recovery actions have the same level of negative impact, if applied in case of a false-positive detection. Thus, it should be considered to define different thresholds for different response and recovery actions. Intuitively, to apply more severe response and recovery actions, the confidence that a certain type of attack has occurred should be higher than for a low-impact response action. This can be expressed by assigning a higher threshold to the actions with severe impact on the general infrastructure.

In order to define suitable thresholds, several factors need to be taken into account. These include the risk of applying the respective response and recovery action, i.e., the damage caused by applying the response and recovery action in case of a false-positive detection, and the risk of not applying the mitigation action in case of a true-positive detection. Modeling the threshold for a specific kind of attack can hence be considered as finding a trade-off between these two aspects. At the same time, the defined thresholds need to be in the range of the used certainty metrics. Note that it might be necessary to have requirements additional to the certainty threshold to trigger certain response and recovery actions.

An example of a low-risk response action is to put a machine on a watchlist. In the case of a false-positive detection, this machine could simply be removed from the watchlist. Since neither putting a machine on a watchlist, nor deleting it from the list is associated with high cost, the damage caused by a false detection is low. If this action is triggered too often, however, it might drastically increase the number of machines on the watchlist, which is probably not desirable.
As a medium-risk response action, the notification of the local IT-admin could be considered. This could allow the local IT-staff to prepare for the case that more severe actions might be taken in the future. It could thus allow to set up backup machines to decrease downtime, if a potentially infected machine is isolated from the network later on. Sending an automatic notification is not an expensive operation, and does not cause direct damage (e.g., in a financial sense) in case of a false-positive detection. But it should be assured that not too many false-positive detections trigger this operation, since otherwise the local IT-staff could tend to ignore these warnings in general, rendering this response action useless.

The automatic isolation of a machine from the rest of the network can be considered as an example for a high-risk response action. Since such an action might be associated with notable financial cost (e.g., if the machine in question is a server hosting an online shop), it should only be taken if there is a negligible chance of the detection being a false-positive.
In general, the set of automated response and recovery actions should be chosen carefully to avoid significant damage caused by the automated system. It is preferable to automate response and recovery actions that are applied in large quantities to reduce the workload of the CSIRT, and that are reversible without causing notable damage.

It could also be considered to model dynamic thresholds that are adjusted automatically. For example, the thresholds could be adjusted according to the false-positive detection rate of the past X-days (sliding window). This approach allows to tackle the problem of new variations of an attack, since these might cause the false-positive rate of the detection system to increase until the detection system is updated to also detect this new variation of the attack reliably. These considerations are at this stage mostly of theoretical nature, since in practice it needs to be assured that the thresholds are adjusted without reaching infeasible levels of confidence. Both too low and too high thresholds would cause damage in practice.

5.3 The Problem of Conflicting Detections

So far, we have only considered aspects of true-positive confidence and response and recovery actions for single types of attacks. If these are combined, new problems arise. One of these problems is caused by conflicting detections: If two different types of attacks are detected on the same machine with comparable levels of confidence, it might be that the response actions for each of these attacks are associated with comparably low risk, but that the impact of the combination of these response and recovery actions might be significantly more severe.

To counter this problem, a prioritization of response actions could be considered. In this context, a similarity metric expressing similarity to past incidents could be used to break ties in confidence. Another approach would be to notify a human operator and to not automate any response and recovery actions for more complex combinations of attacks.

6 Identification of Scenarios with Moderate Risk

Two of the four showcases defined in WP3 allow to apply response and recovery actions that have a relatively small potential of causing severe damage, and can thus be seen as scenarios with moderate-risk response and recovery actions. In the following, these two showcases are briefly presented and response actions suitable for automation are being discussed. For a detailed description of these showcases, please refer to deliverables D3.1 (Data Selection and Preparation) and D3.4 (Algorithms for Analysis of Cybersecurity Data - Initial Version).

6.1 Showcase 1 (Identification of Phishing URLs)

The first showcase defined in WP3 deals with the identification of phishing websites via binary classifiers, based on either URLs or certificates as input data. The binary classifier based on URLs constitutes the main result of Showcase 1 and classifies a given URL as the malicious (URL directs to a phishing website), or benign. Automated response and recovery actions in case of a detection could include the following:

- Creation of an entry in a "phish log" (automated gathering of evidence that could be used later on)
- Blocking of the request and redirection to a page informing the user that he was about to connect to a phishing website
    - It could be considered to have a button on this page that allows the user to connect to the website anyway, to reduce the damage caused by a false-positive detection (e.g., if the website is crucial for the work of the user)
- Sending an automatic informative email/notification to the user associated with the account/machine
- For very high confidence, an automatic password reset could be considered, if the connection request has not been blocked

6.2 Showcase 2 (Detection of DGA Activity)

The second showcase defined in WP3 deals with the detection of Domain Generation Algorithm (DGA) activity. DGAs generate domain names and are commonly used by malware to enable communication with the attacker:

- The malware instance on the victim machine generates a set of domains and tries to connect to them
- The attacker also generates domains and sets up command-and-control servers under these domains
- Because of the properties of DGAs, there is a high chance that one of the domains generated by the attacker is in the set of domains generated by the malware instance
- Since these DGA-generated command-and-control domains are only used for very short intervals, it is very difficult for authorities to take them down in time

In Showcase 2, a binary machine learning classifier is trained to determine, whether a domain has been generated by a DGA. Additionally, a multi-class classifier is trained that can additionally assign a DGA-generated domain to the DGA-family used for its generation.

Response and recovery actions that are suitable for the detection of DGA activity could include the following:

- Putting a machine on a watchlist (unusually high amount of DGA activities can indicate a malware infection)
- Trigger malware scan on the machine

- Notifying the local IT-admin
    - Information that the machine might be infected
    - Allows the local IT-administration to
        - Manually take a look at the machine
        - Set up backup machines in case of an actual infection to decrease downtime
- For high confidence: Isolation of the machine from the network to prevent spreading of malware and its communication with the command-and-control server

Especially for Showcase 2, it might be beneficial to consider sequences of detections for individual different machines. If a machine only tries to connect once to a domain that the classifiers identifies as being malicious, it is comparably unlikely that the machine is infected and that the DNS request was of malicious origin. If, on the other hand, a machine tries to connect to a large number of domains that are identified as DGA-generated, perhaps even in regular intervals, it seems more likely that the origin of the requests is a malware instance that tries to connect to a command-and-control server. By considering the history of a machine, even the result of a binary classifier can be used to reach different levels of confidence.

7 Conclusion

In this preliminary version, we provide the necessary background and related works in the research direction of automating response and recovery steps for cybersecurity incidents. As part of this deliverable D4.6, we progress towards designing a general framework that can capture the approaches and algorithms for automating some types of responses actions depending on several factors. We identify two showcases from WP3 (phishing and DGA) that have moderate response risks and discuss the steps of automation in line with our proposed framework. Further development and concrete evaluation will be part of the final deliverable D4.7 of the task T4.4.

8 References

1. Shahid Anwar, Jasni Mohamad Zain, Mohamad Fadli Zolkipli, Zakira Inayat, Suleman Khan, Bokolo Anthony and Victor Chang. From Intrusion Detection to an Intrusion Response System: Fundamentals, Requirements, and Future Directions, in Algorithms 10, 39, 2017.
2. Marc Burian. Machine Learning based Handling of Cyber Security Incidents, Master Thesis, RWTH Aachen, 2019
3. J. Creasey, Cyber Security Incident Response Guide, 2013.
4. B. Heitmann, An Open Framework for Multi-Source, Cross-domain, Personalisation with Semantic Interest Graphs. Ph.D. Thesis, National University of Ireland, Galway, 2014.

5. J. Bobadilla, Recommender Systems Survey, Knowledge-Based Systems 46, 109–132, 2013
6. ISO/IEC, ISO/IEC 27001 Information Security Standard; Information technology - Security techniques - Information security - Information security management systems - Requirements, 2017.
7. https://at.cis-cert.com/News-Presse/Newsletter/NL-Dez-2016/ISO-Survey-2015.aspx
8. N. Stakhannova, S.Basu, J. Wong, A taxonomy of intrusion response systems, Int. J. Information and Computer Security, Vol. 1, No. 1/2, 2007
9. M. Bromiley, "The Show Must Go On! The 2017 SANS Incident Response Survey“, SANS Institute Information Security Reading Room, 2017.
10. Z. Inayat, A. Gani, N. B. Anuar, M. K. Khan, S. Anwar, Intrusion response systems: Foundations, design, and challenges, Journal of Network and Computer Applications 62, 53–74, 2016
11. IBM Cloud Event Management, https://www.ibm.com/support/knowledgecenter/SSURRN/com.ibm.cem.doc/em_prioritize.html (Accessed: 20.07.2020)
12. Carnegie Mellon - Information Security Office, Computer Security Incident Response Plan. https://www.cmu.edu/iso/governance/procedures/docs/incidenresponseplan1.0.pdf. (Accessed: 20.07.2020)
13. P. Cichonski, T. Millar, T. Grance, K. Scarfone, Computer Security Incident Handling Guide: Recommendations of the National Institute of Standards and Technology. [Online] https://nvlpubs.nist.gov/nistpubs/specialpublications/nist.sp.800-61r2.pdf. (Accessed: 20.07.2020)
14. J. Kick, Cyber exercise playbook, Mitre Corp Bedford, 2014.
15. C. Alberts, A. Dorofee, G. Killcrece, R. Ruefle and M. Zajicek, Defining Incident Management Processes for CSIRTs, Pittsburgh, PA 15213-3890: Carnegie Mellon Software Engineering Institute, 2004.
16. P. Kral, The Incident Handlers Handbook, SANS Institute, 2011.
17. B. E. Strom, J. A. Battaglia, M. S. Kemmerer, W. Kupersanin, D. P. Miller, C. Wampler, S. M. Whitley and R. D. Wolf, Finding Cyber Threats with ATT&CK-Based Analytics, Annapolis Junction, MD: The MITRE Corporation, 2017.
18. https://cve.mitre.org/index.html (Accessed: 20.07.2020)
19. https://www.circl.lu/ (Accessed: 20.07.2020)
20. https://www.auscert.org.au/ (Accessed: 20.07.2020)
21. https://us-cert.cisa.gov/ncas/current-activity (Accessed: 20.07.2020)
22. https://www.mycert.org.my/ (Accessed: 20.07.2020)
23. https://www.misp-project.org/ (Accessed: 20.07.2020)