Risk-Based Safety Scoping of Adversary-Centric Security Testing on Operational Technology

Alexander Staves[a,b,*], Antonios Gouglidis[b], Sam Maesschalck[c,b], David Hutchison[b]

[a]Bridewell, 40 Caversham Rd, Reading, RG1 7EB, United Kingdom
[b]Lancaster University, InfoLab21, Lancaster, LA1 4WA, United Kingdom
[c]RHEA Group, ESEC, Place de l'ESA 1, 6890 Redu, Belgium

Abstract

Due to the recent increase in cyber attacks targeting Critical National Infrastructure, governments and organisations alike have invested considerably into improving the security of their underlying infrastructure, commonly known as Operational Technology (OT). The use of adversary-centric security tests such as vulnerability assessments, penetration tests and red team engagements has gained significant traction due to these engagements' goal to emulate threat actors in preparation for genuine cyber attacks. Challenges arise, however, when performing security tests on these as the nature of OT results in additional safety and operational risk needing to be considered. This paper proposes a framework for incorporating the assessment of safety and operational risks within an overall scoping methodology for adversary-centric security testing in OT environments. Within this framework, we also propose a hybrid testing model derived from the Purdue Enterprise Reference Architecture and the Defense in Depth model to identify and quantify safety and operational risk at a per-layer level, separating high and low-risk layers and being subsequently used for defining rules of engagement. As a result, this framework can aid vendors and clients in appropriately scoping adversary-centric security tests so that depth-of-testing is maximised while minimising the risk to safety and to the operational process. The framework is then evaluated through a qualitative study involving industry experts, confirming the framework's validity for implementation in practice.

Keywords: ICS, CNI, OT, Cyber Security, Security Testing, Scoping

1. Introduction

In the past decade, cyber attacks targeting Critical National Infrastructure (CNI) have risen dramatically [1]. While government bodies and organisations across all 13 sectors of CNI [2] have invested considerably into preparation for cyber incidents, the use of cyber-warfare as part of military strategy in modern-era wars [3] has demonstrated an additional need for securing CNI against threat actors.

As part of the risk assessment process for preparation against such cyber incidents [4], adversary-centric security testing is defined as the use of adversary-based techniques to emulate the actions of a cyber adversary within a defined environment. Engagements categorised as adversary-centric security tests include vulnerability assessments, penetration tests, red team engagements, and more. These tests present three main benefits for asset owners: first, they identify existing vulnerabilities to be patched. Secondly, they assess existing defensive measures such as firewalls or Intrusion Detection Systems to determine and improve current capabilities. Lastly, for advanced tests such as red team engagements, they assess and train Cyber Incident Response Teams in preparation for genuine cyber incidents.

While the use of adversary-centric security tests has been wildly adopted for testing traditional Information Technology (IT) environments [5], many technical challenges that exist within Operational Technology (OT) environments, commonly used within CNI, make it challenging to perform these in a similar manner [6]. One such challenge is the additional risk present in industrial environments due to Industrial Control Systems (ICS) being used to control a physical process. If a technique employed during a security test results in an ICS

*Corresponding author

Email addresses: alexander.staves@bridewell.com (Alexander Staves), a.gouglidis@lancaster.ac.uk (Antonios Gouglidis), s.maesschalck@rheagroup.com (Sam Maesschalck), d.hutchison@lancaster.ac.uk (David Hutchison)

having reduced availability or entering an error state, this effect could change the physical process and bring about a severe loss of safety and a danger to life [7]. Furthermore, legacy OT equipment being designed for environmental resilience over performance is generally poorly equipped for additional overhead caused by these testing tools.

In recent years, newer product lines from OT vendors, such as Siemens [8] or Allen-Bradley [9], have seen an increase in performance, allowing for more flexibility during adversary-centric security testing. However, identifying and understanding the risk that tools and techniques used during such engagements still needs to be undertaken so that scoping of such tests can consider these risks to not disrupt the operational process. This paper provides a methodology for identification and quantification of safety and operational risk during security testing and proposes a framework to scope adversary-centric security tests as a means of maximising the depth-of-testing while minimising safety and operational risk. The core contributions of this paper are:

- A methodology for identification of hazards and deduction of sub-hazards during an adversary-centric security test.
- A methodology for quantifying the safety and operational risk of events that can be caused during an adversary-centric security test.
- A model for taking safety and operational risk into consideration when scoping adversary-centric security tests.
- A framework for aiding test providers and asset owners in scoping safety-risk-aware adversary-centric security tests.

The remainder of this paper is structured as follows. Section 2 provides a background and analysis of related work. Section 3 covers the research methodologies used in the paper. Section 4 details the methodology for identifying safety and operational risks of adversary-centric security test on OT. Section 5 details the methodology for assessing and quantifying the safety and operational risks that are identified in Section 4. Section 6 introduces our framework for scoping of adversary-centric security tests alongside the model used for it. Section 7 covers the design of the qualitative study with participants involved in ICS/OT adversary-centric security testing that we use for the evaluation of the framework. Section 8 reiterates the design and purpose of the framework to lead into a discussion of the results from the quantitative testbed evaluation and the qualitative semi-structured interviews. Section 9 concludes the paper and presents suggestions for future work.

2. Related Work

Due to the high risk associated with causing additional overhead within an ICS/OT environment, such as through tools or techniques employed during active adversary-centric security testing, the majority of research conducted for security testing has been through moving the environment being tested away from the live environment [10, 11, 12] or development of specialised tools [13, 14, 15] for ICS/OT.

As a means of performing risk avoidance for testing of ICS, a majority of research on ICS/OT security has focused on the construction of physical testbed or digital twins. Green et al. propose a model for the design of ICS testbeds for this purpose [10]. Similarly, Gardiner et al. describe their lessons learnt from building an ICS and Industrial Internet of Things testbed [11]. The methodologies described in both of these papers provide a starting point for good practices when designing and developing ICS testbeds for security research. While these testbeds can be used to identify device-specific vulnerabilities and discover ICS-based zero-days [16, 17], their generally lower-scale representation of live environments is better suited for host-level testing. Therefore, it is difficult to assess the full extent of an entire environment's security posture due to the many interactions between large groups of devices. However, one advantage of using ICS testbeds is that they can aid in determining the resilience of specific OT devices against tools and techniques that are planned to be used prior to an adversary-centric security test. This can be used to assess the risk these tools and techniques pose to a live environment without directly interacting with it. Similarly, digital twins, such as the one proposed by Dietz et al., for integration within Security Operations Centers [12] can also be used for similar purposes. However, while their virtual nature reduces the cost of development and increases the flexibility of implementation, they are generally less equipped for vulnerability research and instead used for simulations or direct monitoring.

Several specialised tools such as SimaticScan and PLCScan have been developed as part of an initiative to perform safe and efficient scans on ICS. PLCscan, for example, developed by Dmitry Efanov, is a tool written in python that is able to scan PLCs through Modbus or S7COMM [14]. This tool can query a range of data from the target PLC such as module name, firmware version, PLC name, serial number and more. However, no other functionality is possible; therefore, further assessment would need to be done manually or using other tools. Antrobus et al. identified the limitations of PLCScan and built upon it by proposing a Proof of Concept for SimaticScan [13]. The authors note that SimaticScan goes "beyond simply identifying potential vulnerabilities to verifying the existence of these vulnerabilities" for the target PLC. This is done through three distinct phases: reconnaissance scans, vulnerability assessment and fuzzing. The reconnaissance scan's functionality is similar to that of PLCScan in retrieving the PLC's information for CVE query alongside an SNMP scan. After this, SimaticScan can analyse PCAP files to identify session IDs and plaintext vulnerabilities, perform a dictionary attack on any identified web server login forms, simulate a DoS attack on the PLC, simulate TCP hijacking, and verify unauthorised read-/write access to the PLC Data Blocks. Finally, the tool can fuzz a PLC to determine other vulnerabilities. Overall, while the depth-of-testing of SimaticScan is extensive, its use is restricted to testing of Siemens devices only, severely limiting its effectiveness in environments that deploy devices from multiple vendors. As a means of aiding asset owners in selecting the appropriate tools for their environment and needs, Samanis et al. developed a taxonomy for contrasting ICS Asset Discovery Tools [15], which includes PLCScan. The taxonomy categorises the selected tools into three main classes: Specification, Execution and Output. Specifications of the tool detail its mode of operation, license scheme, scope, and supported protocols. The Execution category describes the tool's method of operation, its usage methodology, user interactivity, and approach to scanning. Finally, the Output category describes the tool's output, such as listening ports, service identification, device info, deployment-specific information and vulnerability identification. Throughout this research, the authors note that none of these tools has information concerning their effect on the operational process; highlighting the need to perform a safety risk assessment prior to their utilisation. However, no methodology is provided for assessing each tool's risk to the operational process when being used as part of an adversary-centric security test.

Existing research has covered model-based testing to provide guidance, but there has been little work related to OT environments. A 2012 paper [18] offers a survey of MBST techniques and related models, highlighting new methods and tools under development in the European ITEA2 project DIAMONDS. Key areas within MBST include security functional testing, model-based fuzzing, risk- and threat-oriented testing, and the application of security test patterns. Work has also been done to highlight the interplay between safety and security in OT, and [19] introduces a risk evaluation methodology to prioritize and manage identified threats, taking into account the inter-dependencies between security and safety. This methodology uses industry-standard metrics such as the Common Vulnerability Scoring System, Security Level from IEC 62443, and Safety Integrity Level from IEC 61508. The authors also emphasize the importance of understanding and treating risks arising from the interplay of safety and security in industrial environments. This is further strengthened by further work in this area [20] which highlights the absence of a holistic approach in existing standards to address conflicts between these domains, such as the potential clash between security authentication and immediate access to safety functions. The need for analysing the risk of security activities and safety is therefore clearly important in this area.

This existing risk is further confirmed by previous work [6], which identified that the safety-critical nature of ICS/OT environments requires unique scoping of adversary-centric security tests so that safety risks can be minimised while ensuring that depth-of-testing is maximised. While testing multiple OT devices from vendors, including Siemens and Allen-Bradley, we identified two main factors of using adversary-centric security testing tools that could cause a reduction in availability or integrity and disrupt the operation process. Firstly, high network traffic generated by these tools could cause an increase in latency or an observable loss in transmitted packets. Secondly, the data being sent to the target could cause additional overhead on its resources, resulting in either a reduction in availability through resource exhaustion or total loss of availability due to some data not being processed appropriately and causing a system crash. While some of the tools used during testing consistently resulted in a severe loss of availability, this loss of availability was not the case for a majority of them; demonstrating that adversary-centric security testing within ICS/OT environments is indeed possible if the effects of the tools and techniques used are understood and taken into consideration during scoping of engagements. The following sections, therefore, provide a methodology for identifying and quantifying the safety and operational risks of conducting adversary-centric security tests within ICS/OT environments and how this security test can aid in scoping these.

3. Summary of Research Methodology

The methodology presented in this paper is a comprehensive approach that combines theoretical exploration with empirical validation. It aims to develop a framework for risk-based scoping of adversary-centric security tests within ICS/OT environments. Our investigation begins with an extensive review of existing literature, establishing a foundation of the current security testing practices and the risks associated with conducting these engagements in ICS/OT environments.

We then devise a research strategy that employs both deductive and inductive reasoning. The deductive elements are rooted in the application of existing safety and risk assessment methodologies, such as HAZOP and FTA, to the domain of ICS/OT security testing. This approach provides a structured, hypothesis-driven exploration of potential risks, which are then empirically examined through a series of controlled testbed experiments. These experiments simulate adversary actions within OT systems, providing data on the impact of various security testing techniques on system integrity and operational continuity.

To complement the quantitative data from these experiments, we conduct a qualitative study involving semi-structured interviews with a carefully selected group of industry experts. This exercise allows for the validation and refinement of our findings through the perspectives of experienced professionals in the field. The experts are chosen for their demonstrated expertise and diverse views on OT security, ensuring a holistic understanding of the risks and challenges in conducting security tests within these environments.

The integration of quantitative and qualitative data is crucial in developing and evaluating our risk-based scoping framework. It ensures that the framework is not only based on empirical evidence but also aligns with the practical realities of OT security testing, as observed by industry practitioners.

The result of this methodological approach illustrated and summarised in Figure 1, leads to the creation and optimisation of a risk-based framework for scoping adversary-centric security tests, as presented in this paper. The framework is designed to guide test providers and asset owners in systematically evaluating and mitigating safety and operational risks during security testing in OT environments. The following section starts the investigation into the identification of safety and operational risks, which are crucial to take into account for adversary-centric security tests.

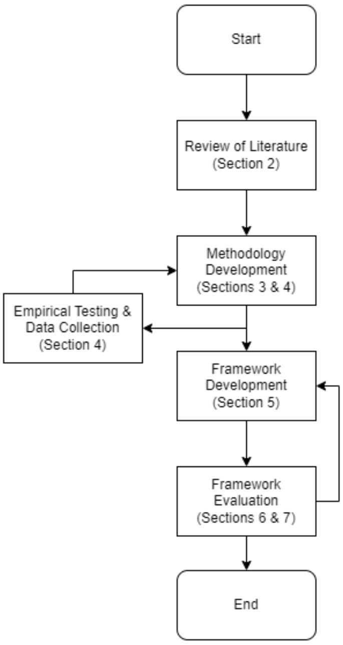

Figure 1: Summary of Research Methodology

4. Identifying Safety and Operational Risks of Adversary-Centric Security Testing on ICS/OT

4.1. Identifying hazards with (C)HAZOP

Derived from the well-established Hazard and Operability (HAZOP) study [21], a Control Hazard and Operability (CHAZOP) study provides a comprehensive framework for reviewing controllability, safety and operability issues during the implementation of ICS/OT [22]. The objective of such a study is to understand and assess hazards that could cause a loss of safety or a disruption to the operational process, which is the first step in quantifying the safety and operational risks of conducting adversary-centric security tests within ICS/OT environments. While several methodologies exist for identifying hazards, HAZOP was found suitable for the identification of hazards caused by adversary-centric security tests within ICS/OT environments due to its applicability for identifying both safety and operational hazards and its widespread application across several domains, including manufacturing, engineering and CNI [23]. Additionally, while the Institution of Chemical Engineers acknowledge that certain factors such as no prior design review; inappropriate, incompetent or too many team members; lack of operational experience; defensive designers; and arrogant project managers can reduce the effectiveness of (C)HAZOP studies, if executed correctly, these type of studies allow for effective and cost-efficient qualitative risk assessment [24]. While (C)HAZOP is not a new methodology for identifying operational and safety hazards, its application to adversary-centric security testing in ICS/OT environments is innovative. By using it in this context, we can identify the risks associated with conducting these engagements. This, in turn, guides efforts to reduce or mitigate these risks, ensuring that adversary-centric security tests are conducted safely.

When applying these studies to adversary-centric security testing, the terminology used is similar to that used in HAZOP studies with additional context. These are as follows:

- Node: The specific location in the process for which deviations can occur (for example: heater, liquid tank, mixers).
- Parameter: The parameter for the condition(s) of the process (for example: temperature, level, flow, pressure).
- Intent: How the node is designed to operate under normal conditions.
- Guidewords: Terms when considered with one or more parameters that form a hypothetical deviation for risk consideration (i.e GUIDEWORD + PARAMETER = DEVIATION).
- Deviations: Events that lead to a partial or total disruption of the operational process.
- Causes: The combination of the events that cause deviation.
- Consequences: The outcome derived from the causes that could lead to operational impact or loss of safety.
- Actions: Actions that can be taken to mitigate the identified risk(s).

The methodology for applying a (C)HAZOP study in the context of adversary-centric security testing is depicted in Figure 2. As opposed to HAZOP, (C)HAZOP focuses on hardware and software design of ICS/OT rather than vessels and pipes. Any system related to safety or operation functions should be considered during the study. For each of the identified endpoints, the following must be considered to comprehensively understand the risk that these face:

- The functionality of the system.

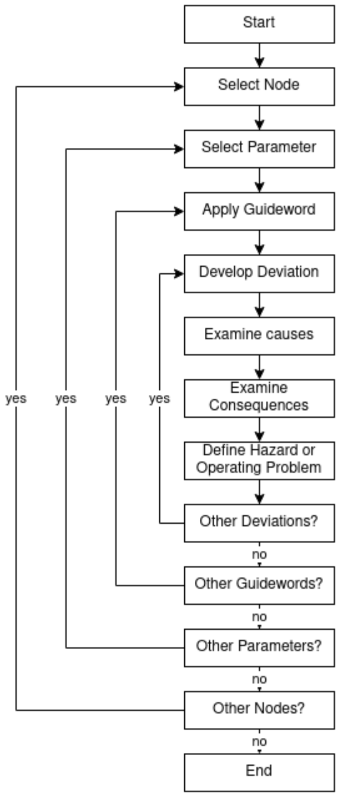

Figure 2: HAZOP Methodology

- All the dependencies of the system.
- Segregation and redundancy deployments.
- Application of the guidewords from Table 1.

While the methodology in Figure 2 is generally only applied to endpoints used for safety-related functions, for adversary-centric security testing, parameters that affect the operational process should also be considered. When applying (C)HAZOP to the risk assessment process, it is essential to ensure the full coverage of documents is considered and should include the following:

- User Requirement Specification and Detailed Functional Specification documents.
- Piping and Instrumentation (P&I) Diagrams.
- Network Diagrams.
- System hardware configuration documents.
- Power and wiring documents.
- Channel/loop diagrams.

Guideword	Definition	Example
NO or NOT	Complete Negation of the Intention	No Flow; No Communication; No Pressure
MORE and LESS	Quantitative Increase or Decrease	More/Less Flow; Less Communication; More/Less Pressure
AS WELL	Qualitative Increase	Intended Valve Close As Well As Unintended Valves
PART OF	Qualitative Decrease	Part Of Intended Valves Closing
REVERSE	Opposite of the intention	Reverse Flow; Reverse Direction
OTHER THAN	Complete Substitution	Other Than X Chemical

Table 1: (C)HAZOP guidewords

- System malfunction fail-safes.

To demonstrate an example application of (C)HAZOP for identifying hazards during an adversary-centric security test, we have applied this process to a scenario engineered within our ICS testbed. This testbed has been developed over the past nine years and emulates real-world scenarios using physical industrial devices such as PLCs and HMIs from various vendors, including Siemens, Schneider Electric, Allen-Bradley and ABB based on [10].

Our scenario consists of an operational process to manually control the water levels of a tank through an HMI panel and contains the following elements:

- Siemens SIMATIC ET-200S (physical device): sends data to the HMI, receives commands from the HMI, receives data from the water tank sensor, and sends commands to the tank pump and release valve.

- Siemens TP1500 Basic PN HMI (physical device): displays water tank levels, receives data from PLC, sends open/close commands for both the tank pump and the release valve of the water tank to the PLC.

- Water Tank (virtualised): Container for water storage.

- Water Tank Pump (virtualised): turns on and off to increase water level in the water tank.

- Water Tank Release Valve (virtualised): opens and closes to decrease water level in the water tank.

- Water Tank Sensor (virtualised): sends water level data to PLC.

Figure 3 represents an ANSI/ISA-5.1-2009 [25] and ISO 14617-6:2002 [26] compliant P&I Diagram of the scenario developed within our ICS testbed. Despite the scenario being simple in concept, it accurately depicts, at a reduced scale, the potential hazards possible within real-world industrial processes. By applying a (C)HAZOP methodology, several safety and process hazards can be identified and are provided in Table 2.

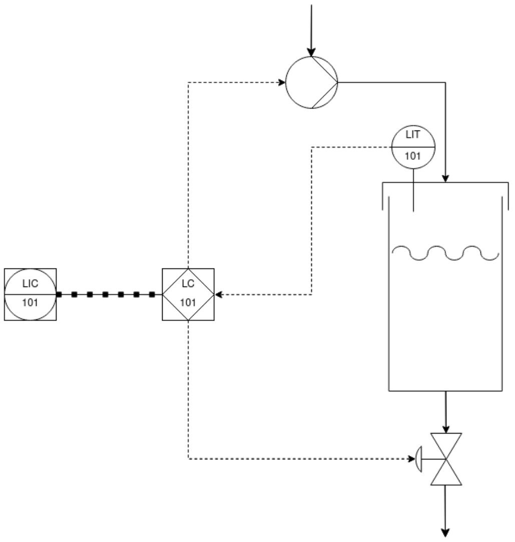

Figure 3: P&I Diagram of Water Tank Scenario

4.2. Establishing Risk Events and Causes with FTA

Following the identification of hazards using (C)HAZOP, a Fault Tree Analysis (FTA) can be conducted to further decompose hazards into their causes. While (C)HAZOP can also be used to qualitatively identify the causes of hazards, FTA is used to provide further depth to this by identifying the relationship between different events that could lead to the cause of a major hazard. This analysis adopts a top-down approach where hazards are broken down into possible causes. Each of these causes is then decomposed until a set of “basic events” is established, for which their risk can be calculated. In the context of adversary-centric security testing, these basic events can be directly associated with specific testing techniques or tools, which in

Parameter	Guideword	Deviation	Causes	Consequences
Tank Water Level	More	More Water Level	Pump On and Water Level High	Tank Overflow
Pump	No	No Pump (De)activation	Pump unresponsive	Disruption to Operational Process
Release Valve	No	No Valve (De)activation	Valve unresponsive	Disruption to Operational Process
PLC	No	No PLC Communication	PLC Resource Overload; PLC Crash	No Control of Pump and Release Valve
PLC	Less/Late	Less/Late PLC Communication	PLC Resource Overload; Network Congestion	Limited Control of Pump and Release Valve
HMI	No	No HMI Communication	HMI Resource Overload; HMI Crash	No Control of PLC, Pump and Release Valve
HMI	Less/Late	Less/Late HMI Communication	HMI Resource Overload: Network Congestion	Limited Control of PLC, Pump and Release Valve

Table 2: (C)HAZOP Output for Water Tank Scenario

turn contributes to identifying how the aforementioned can contribute to operational or safety impact being realised. The components of a Fault-Tree Diagram (FTD) are defined within IEC 61025 [27] and are as follows:

- Gates: Symbols (see Figure 4) showing the logical relationship between a cause and a consequence. Static gates do not depend on the order of occurrence whereas dynamic gates do.
- Events: Symbols (see Figure 4) describing failure states, system states, or events within an even chain.

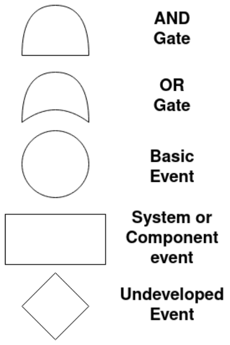

Figure 4: Example Symbols used for Fault Tree Analysis

In order to fully develop a fault tree, a thorough understanding of the cause and effect relationships between a hazard and its subsequent causes is required and can be provided by both safety and ICS engineers. Following a pragmatic methodology, causes need to be determined based on their possibility of occurring during adversary-centric security testing. In contrast with traditional safety risk assessment, we exclude failure mode risks including power failure and mechanical failures, such as the failures of the sensors and actuators, because these cannot be caused by adversary-centric security testing.

Continuing with the example provided by the scenario described in Figure 3, Figure 5 was developed following an FTA analysis for the Tank Overflow hazard that was identified during the preceding (C)HAZOP study. There are of course safety hazard events such as power supply failures or mechanical valve failures that could lead to this hazard scenario and should be developed within a traditional HAZOP study. However, because our study focuses on the effects of adversary-centric security testing on safety and the operational process, such hazards have been excluded from the final FTD.

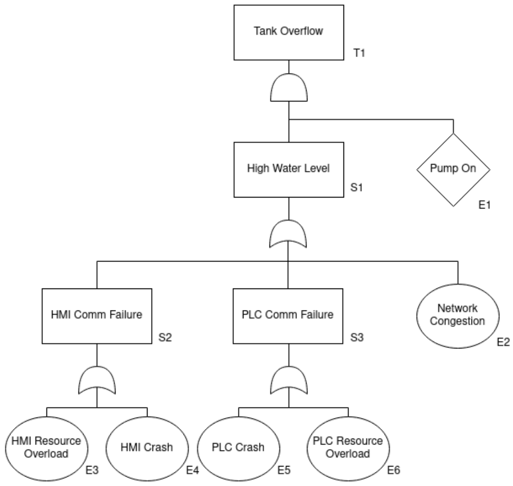

Figure 5: Fault Tree Diagram for Tank Overflow Hazard

Once an FTD has been generated, the minimal cut sets (MCSs) for this can be deduced. These sets are the unique combination of basic events from the FTD that can lead the top event to occur, such as the water tank overflowing from Figure 5. For this, top events are denoted as T, system events are denoted as S, and basic events are denoted as E. When determining the MCSs for an identified hazard, OR gates produce additional cut sets, whereas AND gates make the cut sets more complex. For example, to begin developing the MCSs for the tank overflow scenario, The first AND gate immediately below T1 can be listed as the following expression:

$$S1 \wedge E1 \tag{1}$$

The expansion of $S1 = S2 \vee S3 \vee E2$ leads to the following sets:

$$\begin{aligned} &S2 \wedge E1 \\ &S3 \wedge E1 \\ &E2 \wedge E1 \end{aligned} \tag{2}$$

Substituting for $S2 = E3 \vee E4$; and $S3 = E5 \vee E6$ results in the following MCSs (denoted C_i):

$$\begin{aligned} C_1 &= \{E6, E1\} \\ C_2 &= \{E5, E1\} \\ C_3 &= \{E4, E1\} \\ C_4 &= \{E3, E1\} \\ C_5 &= \{E2, E1\} \end{aligned} \tag{3}$$

While the MCS provided in the list of sets 3 does not require further reduction, more complex cut sets can be reduced by removing redundant events or sets through the idempotence or absorption rule, for example. Because of the complexity of some systems, this can result in MCSs containing several thousand cut sets. Therefore, truncation can be used to remove cut sets that are believed to contribute negligibly to the top event occurring, which can be determined through traditional safety risk assessment. Additionally, if available, FTA software can also be used to automate the creation of Fault Trees and calculation of MCSs.

5. Quantifying Safety and Operational Risks of Adversary-Centric Security Testing on ICS

Once safety and operational hazards have been identified, these can be evaluated to determine the risk of conducting adversary-centric security tests within ICS/OT environments. By understanding and assessing these risks, strategies can be formulated to appropriately scope such engagements and ensure their completeness while mitigating the potential for operational disruption and loss of safety. Safety and operational risk is commonly defined as a product of likelihood and impact, where likelihood refers to the probability of a risk event occurring and impact refers to the severity of the consequences when a risk event occurs. Due to the operational nature of ICS/OT environments, the impact of events can be represented through either monetary cost (for hazards leading to disruption of the operational process) or injuries/deaths (for hazards leading to a loss of safety). Both expert estimation and historical data can be used to calculate the impact of an event occurring in their respective environments. The following subsections describe the methodology for quantifying the likelihood of hazards occurring by calculating the probability of the respective basic events occurring based on safety and operational failures. This qualification can subsequently be used in the overall risk quantification of identified hazards. All data and scripts used for quantification of risk have been made publicly available on GitHub [28].

5.1. Cut Set Probability

As part of the evaluation of a Fault Tree (discussed in section 4.2), the probability of top events can be calculated based on the probability of the bottom events occurring. Because the fault tree of real systems commonly contains recurring basic events, this evaluation can be done using derived MCSs. For example, given the MCSs determined for the scenario described in Figure 3, the top event (Tank Overflow) can be expressed as the following boolean expression:

$$\begin{aligned} \text{TankOverflow} = &(E2 \wedge E1) \vee (E3 \wedge E1) \vee (E4 \wedge E1) \\ &\vee (E5 \wedge E1) \vee (E6 \wedge E1) \end{aligned} \tag{4}$$

As such, the probability for the top event occurring can be expressed as follows:

$$\begin{aligned} P(\text{TankOverflow}) = &P((E2 \wedge E1) \vee (E3 \wedge E1) \vee \\ &(E4 \wedge E1) \vee (E5 \wedge E1) \vee (E6 \wedge E1)) \end{aligned} \tag{5}$$

As each MCS is capable of causing the top event, their likelihood to cause the top event is therefore cumulative. However, each MCS may not be mutually exclusive (i.e. non-disjoint) since these can contain the same basic event. Due to the rule of addition, the probability of each MCS occurring will be greater than or equal to the probability of the top event occurring. For example, E1, E2 and E3 could coincide, satisfying the first two MCSs. Because of this, the upper-bound of the probability of the tank overflowing scenario can be defined as:

$$\begin{aligned} P(\text{TankOverflow}) \leq &P(E2 \wedge E1) + P(E3 \wedge E1) + \\ &P(E4 \wedge E1) + P(E5 \wedge E1) + \\ &P(E6 \wedge E1) \end{aligned} \tag{6}$$

While using term combination does increase the accuracy of the probability of a top event occurring, the resulting formula for this probability becomes exponentially more complex the more MCSs are present, which is especially common for large fault trees. Furthermore, the subsequent combination of terms within a derived formula, otherwise known as the “rare event contribution”, contribute significantly less to the probability of the top event occurring than the first terms established from the FTA. Therefore the approximation provided in equation 6 can be deemed adequately accurate for subsequent risk analysis as it provides an upper bound for the probability of an event occurring.

Because the events contained within an MCS are independent, as per the definition of a basic event, the final upper-bound probability of the tank overflowing can be further decomposed as follows:

$$\begin{aligned} P(\text{TankOverflow}) \leq P(E2) \times P(E1) + P(E3) \times P(E1) \\ + P(E4) \times P(E1) + P(E5) \times P(E1) \\ + P(E6) \times P(E1) \end{aligned} \tag{7}$$

The following formula can, therefore, be used to calculate the upper-bound of the probability of a safety or operational hazard occurring during an adversary-centric security test using MCSs:

$$P(\text{TopEvent}) \leq \sum_{j=1}^{k} \Big[\prod_{E \in C_j} P(E) \Big] \tag{8}$$

Where E is a basic event belonging to a minimal cut set C_j and k is the total amount of MCSs.

5.2. Basic Event Probability

To provide further granularity in determining the risk of top events, the probability of the basic events belonging to the MCSs of an associated top event needs to be calculated. Previous work identified two contributors to basic events leading to safety and operational hazards during adversary-centric security tests [6]. The first is due to the excessive data throughput of tools being used (named Network-Caused Basic Events), and the second is due to the contents of the data. The second event type can be further decomposed into two sub-categories: data that causes excessive overhead (named Resource Exhaustion Basic Events) and data that, when processed by an industrial device, results in a system crash or error (named Incompatible Data Basic Events). From the FTD illustrated in Figure 5 for the water tank scenario, we can determine that basic event E2 can be categorised as a Network-caused basic event, basic events E3 and E6 can be categorised as resource-exhaustion basic events, and basic events E4 and E5 can be categorised as Incompatible data basic events.

5.2.1. Network-Caused Basic Events

To obtain accurate data on how the data throughput of tools or techniques used during an adversary-centric security test on ICS/OT could have an adverse effect on the operational process of an industrial environment, a network stress test can be performed on the target endpoints; done within a testing environment such as a testbed to prevent impact to the operational process. By gradually increasing the amount of data being sent to the target, the endpoint’s capability of responding to high network traffic can be assessed and thus, the throughput of data at which an increase in latency or a drop in packets would lead to disruption to the operational process, can be determined.

Continuing with the example provided in the P&I Diagram from Figure 3, both the Siemens HMI and PLC need to be tested to determine the limits of their packet buffer and the effect of high throughput tools and techniques on these. For this evaluation, a custom script was created to simulate network traffic using ICMP ping packets with decreasing delay between packets to determine the behaviour of these devices with different network throughputs. We should note that flooding active services (e.g. HTTP) might lead to different results. The results from this test can be found in Figure 6 and Figure 7. For the ET-200S, a considerable increase in latency can be observed at around 400 packets per second, equating to approximately 25.6 KB/s (due to each packet used during the test being 64 bytes in size). However, no packet loss is observed until around 40000 packets per second, which equates to a throughput of approximately 2.56 MB/s. However, results from testing the HMI show a near-total packet loss at 1000 packets per second with no increase in latency prior to this. From this data, we can conclude that to prevent any disruption to the operational process, all tools and techniques used during an adversary-centric security test within this environment must not exceed approximately 25.6 KB/s.

While it is possible to determine a maximum tolerable throughput for adversary-centric security testing activities, inherent network jitter can contribute to additional risk in environments with strict timeliness requirements, such as CNI, and therefore must be determined. While some tools might seem safe for use within certain environments due to their low inherent network throughput, they may cause additional jitter leading to the possibility of reduced availability and therefore must also be considered. Several works have attempted to estimate the distribution of network jitter with varying results. For example, Karakas determined that network

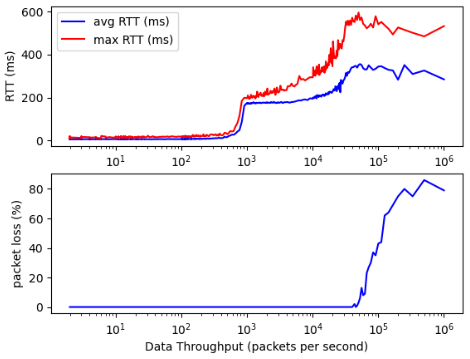

Figure 6: ET200S data throughput test results

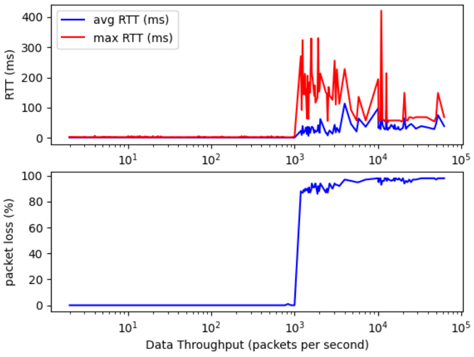

Figure 7: Siemens HMI data throughput test results

jitter distribution can mostly be fitted to a lognormal distribution if no additional factors, such as firewall ruling, contribute to network delay [29]. However, Mozhaiev et al. claim that random jitter is best fitted to a Gaussian Distribution [30] and Daniel et al. describe network jitter as fitting a Laplacian distribution [31]. This disparity in distribution fits is mainly attributed to the causes of network jitter, such as random noise, crosstalk from signals, the effect of dispersion from signal propagation, or resistance mismatch, which all affect the distribution of network jitter differently. As such, the distribution of network jitter is dependent on the environment itself and therefore needs to be determined for each environment which is planned to be tested.

To determine the network jitter distribution within our water tank scenario, we collected data on the latency of both the ET-200S and the HMI while these were continuously receiving 25.6 KB/s of data for 15 minutes. The results of this experiment can be found in figures 8 and 9. Using the python library `distfit`, we calculated the Residual Sum of Squares (RSS) for the best fitting distributions, which were the lognormal distribution (RSS=0.048644 for the ET-200S data and RSS=2.496344 for the HMI data) and the Generalised Extreme Value (GEV) distribution (RSS=0.04456 for the ET-200S data and RSS=2.646364 for the HMI data). All results from using `distfit` to calculate the various fitness scores of distributions can be found in Appendix A. While the GEV distribution was deemed to be a better fit for the jitter distribution of the ET-200S, the shape parameter of the GEV distribution for the HMI was negative, suggesting that this distribution has an upper limit as per its definition when using negative shape parameters. As network jitter can cause, in extreme cases, high latency values leading to packet loss, the GEV distribution was therefore rejected, and the lognormal distribution was selected as the most appropriate distribution fit for the water tank scenario's network jitter.

As such, the following 3-parameter formula can be used to calculate the probability density function for an endpoint's latency based on jitter within the water tank scenario described in Figure 3; the curve for these is illustrated in figures 8 and 9:

$$f(x;m;s;\theta) = \frac{1}{(x-\theta)s\sqrt{2\pi}} exp\left(-\frac{(\ln(\frac{x-\theta}{m}))^2}{2s^2}\right) \qquad (9)$$

$$x > \theta; m, s > 0$$

where:

- x is a given RTT in milliseconds.
- θ is the location parameter of the distribution.
- m is the scale parameter of the distribution.
- s is the shape parameter of the distribution.

It is worth noting that the probability density function of both figures 8 and 9 are represented through a histogram, meaning that the probability for a given RTT range is defined as the following:

$$P(bar_{min} < X < bar_{max}) = (bar_{max} - bar_{min}) \times bar_{height} \qquad (10)$$

Using the parameters derived from either interpolation techniques such as curve fitting or their respective formulas, the cumulative distribution function (CDF) of an endpoint's latency can be used to determine the probability of an adversary-centric security tool or technique's throughput causing undesirable latency and affecting the operational process.

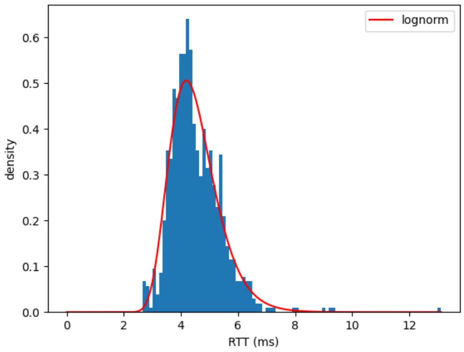

Figure 8: Lognormal distribution curve fit of ET-200S network jitter

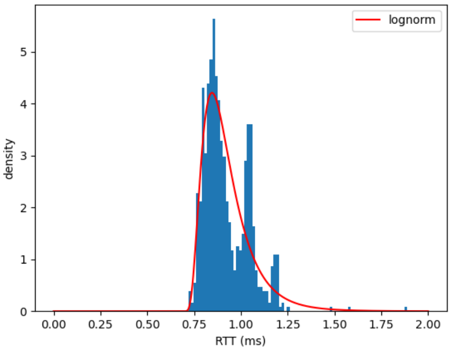

Figure 9: Lognormal distribution curve fit of Siemens HMI network jitter

$$F_x(x; m, s, \theta) = \Phi\left(\frac{ln(\frac{x-\theta}{m})}{s}\right) \tag{11}$$

$$x \geq \theta; m, s > 0$$

where:

- $\Phi(x)$ is the cumulative distribution function of the standard normal distribution ($\Phi(x) = \int_{-\infty}^{x} \frac{exp(\frac{-x^2}{2})}{\sqrt{2\pi}})$

Using the CDF derived from data for the Siemens HMI and the ET-200S, we can estimate the probability of the latency exceeding a tolerable value which must be determined, specific to the environment and endpoints being tested, by safety and ICS engineers. Example tolerable latency values for the HMI and PLC from the scenario described in Figure 3 were arbitrarily determined to be 3ms and 15ms, respectively. The probabilities of the latency of these endpoints exceeding these values while receiving 25.6KB/s of data from adversary-centric security testing tools and techniques were determined as follows using the CDF from equation 11:

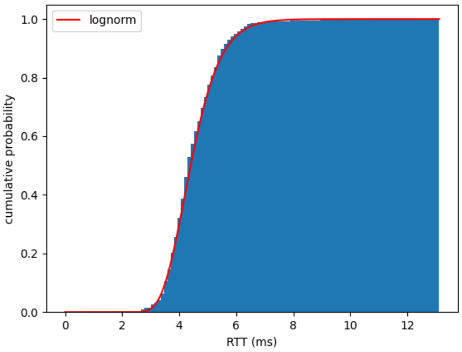

Figure 10: Jitter cumulative probability for ET-200S during 400 packets (of 64 Bytes) per second test

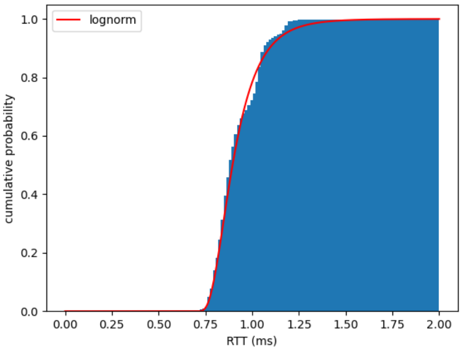

Figure 11: Jitter cumulative probability for Siemens HMI during 400 packets (of 64 Bytes) per second test

$$P(\text{HMI\_latency} > 3ms) = 3.25 \times 10^{-7} \tag{12}$$

$$P(\text{PLC\_latency} > 15ms) = 1.06 \times 10^{-8} \tag{13}$$

These probabilities can subsequently be used in determining the probability of Network-Cause Basic Events that contribute to a top event occurring in the scenario described in Figure 3 as discussed in Section 5.1.

5.2.2. *Resource Exhaustion Basic Events*

While data throughput exceeding tolerable ranges can disrupt the operational process, some tools and techniques employed during an adversary-centric security test may cause similar disruption, due to endpoint resource exhaustion, without exceeding these ranges. Because of this, testing also needs to be done to determine if any tools or techniques planned to be employed throughout an adversary-centric security test could cause disruption due to resource exhaustion. If data from previous engagements is unavailable, this data needs to be obtained through experimentation in a testing environment such as a testbed. Expert opinion can aid in estimating the effect of tools and techniques; however, testing is required for tools or techniques that have an unknown effect on an endpoint's resources.

For example, port scanning is a commonly-used technique employed during adversary-centric security tests to discover open ports on an endpoint. By identifying these, the devices' services can be deduced and tested further for vulnerabilities. Nmap is a popular tool used for port-scanning, allowing for different scan techniques. As such, a comprehensive test of all these techniques must be done to determine which of these presents the least risk, if any, for port-scanning ICS/OT. The following port scan techniques were therefore tested on the ET-200S within the context of the water tank scenario described in Figure 3:

- Idle (control test),
- TCP SYN scan (uses SYN packet but does not complete full TCP handshake),
- TCP Connect scan (full TCP handshake),
- UDP scan (only used to determine open UDP ports),
- SCTP INIT scan (uses the SCTP protocol over TCP/UDP),
- TCP NULL scan (no flags set),
- TCP FIN scan (TCP FIN flag set only),
- TCP Xmas scan (TCP FIN, PSH, and URG flags set),
- TCP ACK scan (ACK flag set),
- TCP Window scan (examines TCP Window field of the returned RST packets),
- TCP Maimon scan (TCP FIN and ACK flags set).

By acquiring data on PLC CPU execution time with no additional load, a baseline can be determined to identify abnormally high increases in execution time, which could disrupt normal functions. The results from running these scanning options continuously for 15 minutes on the ET-200S have been summarised into boxplots, illustrated in Figure 12. We have focused on IP-based protocols, and such these results might not be applicable to other non-IP-based communication used with PLCs These boxplots allow us to identify the non-outlier minimum, non-outlier maximum, median, first quartile, third quartile and outliers of CPU execution times for each scan option.

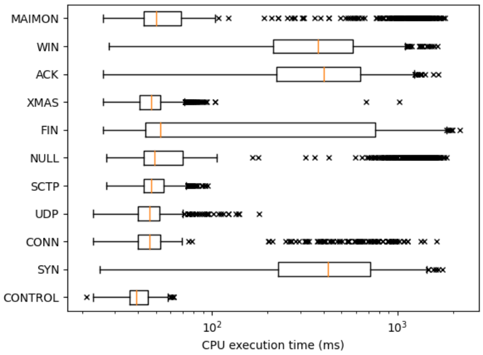

Figure 12: ET-200S CPU execution times with Nmap Scan Options

To obtain additional precision on how these scan options can impact the operational process, data on the effect of these on an endpoint's network response time can be used. As such, Figure 13 summarises the results of testing the ET-200S' latency when being scanned continuously for 15 minutes with the same scan options as Figure 12.

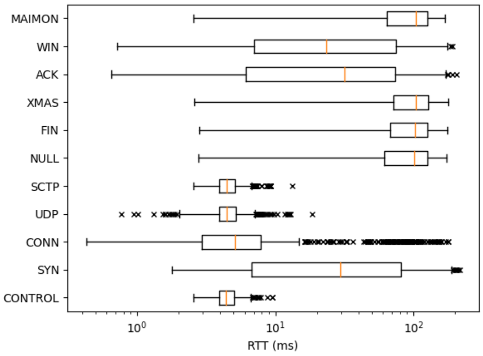

Figure 13: ET-200S Latency with Nmap Scan Options

From this data, four scan options can be identified with certitude as being high risk due to both the observed CPU execution times and latency of these tests being considerably higher than that of the control test. These include the Window, TCP ACK, TCP FIN, and

TCP SYN scans. This higher latency is because these scan types cause additional load on endpoints to increase the scan's stealth or speed. For SYN scans, as an example, the speed of the scan is increased because the TCP three-way handshake remains incomplete. However, due to this, the endpoint continuously allocates resources for incoming TCP connections, which never occur, leading to the potential of a SYN flood and, consequently, the potential to disrupt the operational process.

Despite the Xmas scan, the Maimon scan and the NULL scan not resulting in high CPU execution times, a non-negligible increase in latency was observed. This is most likely the result of these tests being conducted with default scan speeds (no initial scan delay and dynamic parallelism). Therefore, as a means of reducing the risk of latency issues causing disruption to the operational process, less aggressive scan speeds can be used, such as the polite (initial scan delay of 400ms and max parallelism of 1) or sneaky (initial scan delay of 15 000ms and max parallelism of 1) options. However, other scan options, discussed subsequently, with normal scan speeds, present considerably less risk and should be favoured over the Xmas, Maimon and NULL scan options.

While the SCTP scan resulted in a negligible increase in CPU execution time and latency for the ET-200S, the expected scan results were not returned. Despite TCP port 102 (S7COMM) being open, the SCTP scan identified it as closed. This is due to the PLC not supporting this specific protocol and therefore not replying with appropriate data for identifying open ports. This scan option is, therefore, not recommended for use on this PLC specifically.

The remaining scan options, which include the UDP scan and TCP Connect scan, resulted in both a negligible deviation of CPU execution time and acceptable increases in latency while also returning correct information on open ports. The UDP scan is unique because it is the only scan option available for identifying open UDP ports. Fortunately, using this option on the ET-200S does not result in any significant increase in CPU execution time and can therefore be considered safe to use depending on established risk tolerance. In our test environment there were no UDP services running on the targeted PLC. While causing some increase in CPU execution time, the TCP Connect scan causes less disruption than the other tested scan options. However, these increases in CPU execution time are expected as any additional load on the PLC will lead to increased CPU execution time regardless of the task. Furthermore, these outliers (*execution time* $> 100ms$ for the TCP Connect scan) only consist of 9% of total registered execution times. Additionally, the speed of the scan can be configured to reduce increases in latency and further reduce risk to the operational process. We can, therefore, conclude that, if within established tolerable ranges, both the UDP and TCP Connect scans are the safest scan options for use on the ET-200S.

Due to the Siemens HMI not having diagnostic capabilities, acquiring data on resource usage is more challenging than for the PLC. Despite this, measuring network latency alone, while not as accurate as measuring both this and CPU execution time, provides sufficient estimation of the effects of different port scanning options due to changes in latency, in most cases, correlating with resource usage as observed when testing the ET-200S. The results of running these scan options continuously for 15 minutes on the Siemens HMI are summarised in Figure 14.

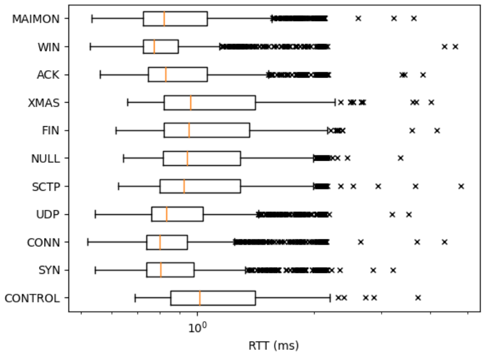

Figure 14: Siemens HMI Latency with Nmap Scan Options

Because the HMI runs WinCC on top of a Windows Operating System (Windows Embedded Compact) and uses better hardware than the PLC, it is considerably more resilient to the different scan options available with Nmap. As seen in Figure 14, all of the scan options used on the HMI returned similar results and no considerable increase in latency was observed as opposed to the results from conducting the same test on the ET-200S. Therefore, we can conclude that most scanning options can be considered safe for use on the HMI. However, if additional risk reduction is required, this can be done by using less aggressive scan speeds, similar to the ET-200S.

5.2.3. Incompatible Data Basic Events

Despite basic events caused by incompatible data presenting the most danger to the operational process, identifying and quantifying these is relatively simple. During testing for resource exhaustion basic events, any tool or technique which consistently results in the failure of data integrity or the failure of exception handling needs to be identified and marked during scoping to prevent the use of these during an engagement.

During testing of the ET-200S used for the water tank scenario described in Figure 3, three open-source and commercial tools were identified as affecting the PLC's behaviour to the point of disrupting the operational process: Nmap (service and version enumeration), Nessus, and OpenVAS. While running these, the PLC would enter an error state, disrupting all communication to the HMI and actuators and requiring both a complete power cycle and a master reset to restore the PLC to a working state.

Despite Nessus and OpenVAS initially causing total disruption to the operational process, a change in the configuration of the ET-200S was identified to prevent the PLC from entering an error state. By loading a programming error Organisation Block (designated OB121 in the TIA Portal, used for programming Siemens PLCs) into the CPU load memory, subsequent scans using Nessus and OpenVAS did not result in any error state occurring. However, further testing following the methodologies described in Sections 5.2.1 and 5.2.2 would still need to be undertaken to determine network and resource-related risks when using these tools.

While loading OB121 into the CPU load memory resolved error states caused by using Nessus and OpenVAS, subsequent scans using Nmap's service and version enumeration module still resulted in the PLC entering an error state. Upon analysis of the packets sent by Nmap prior to the PLC crashing, Nmap attempts an RDP Negotiation Request with the PLC as part of an RDP Connection Request Protocol Data Unit. Due to the PLC's inability to process this request, it enters an error state, disrupting all communication to the HMI and actuators. No solutions were identified for preventing the PLC from entering an error state. Therefore, using Nmap's service and enumeration feature was deemed too high risk and should be categorised as prohibited during scoping of adversary-centric security testing for the ET-200S.

Similarly to the tests performed on the HMI in Section 5.2.2, running Nmap's service and version enumeration option, Nessus, and OpenVAS did not result in any abnormal behaviour; signifying that the use of these tools on the HMI does not cause Incompatible Data Basic Events.

6. Risk-Aware Scoping of ICS/OT Adversary-Centric Security Testing

6.1. Model Proposal for Zone and Level Scoping of Adversary-Centric Security Tests

While scoping of adversary-centric security testing for IT is often client-defined, the existing safety and operational risks discussed and quantified in Section 5, when conducting security tests within ICS/OT environments, provide further constraints for the scoping of these. This scope, therefore, requires further granularity to ensure that no disruption to the operational process is observed. As such, when defining the scope of an adversary-centric security test within industrial environments, a layered methodology can be used to separate the scoping of zones and levels containing differing levels of risk.

To enable this, we propose a hybrid model called the Testing in Depth for ICS (TiDICS) methodology, derived from the Purdue Enterprise Reference Architecture (PERA), also known as the Purdue Model, and the Defense in Depth Model. PERA is a commonly used architecture for segmenting devices and equipment within an ICS/OT environment into hierarchical functions. For the proposed framework model, an extended version of this architecture which utilises a Demilitarized Zone (DMZ) to provide additional separation between the Enterprise and Manufacturing Zones has been selected [32]. Its use can be applied to the scoping of adversary-centric security testing as the different zones and levels within the model, illustrated in Figure 15, have different risk levels due to the different device types implemented in each zone or level. The following provides a short description of the zones and levels within the Purdue Model and the safety and operational risk that exist within each of these:

Enterprise Zone - Levels 4-5:

- Level 5 - Enterprise: This level corresponds to where the centralised IT systems and functions reside. This includes business-to-business, business-to-customer and resource management services. At this level, there is typically no requirement for direct access to industrial equipment. Therefore, testing devices at this level has little-to-no impact on the operational process.

- Level 4 - Site Business Planning and Logistics: This level corresponds to where functions and systems which require access to the Enterprise level (level 5) reside. These functions include enterprise

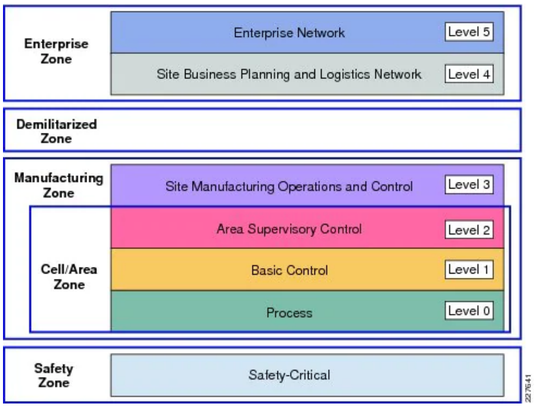

Figure 15: The Extended Purdue Enterprise Reference Architecture [32]

network services such as internet access, e-mail, non-critical production systems, and enterprise applications. While no direct link to the Manufacturing zone is made at this level, resulting in little-to-no safety and operational risk to these during an adversary-centric security test, the open nature of the systems on this level could be used by a threat actor to pivot to lower levels and should be considered during scoping of a security engagement.

Demilitarized Zone:

- The DMZ is used within the extended version of the Purdue Model to provide a segregation zone between the enterprise zone and manufacturing zone of an overall network. This allows for effective segmentation of organisational control by preventing direct communication between these. Adversary-centric security tests can, therefore, be used to ensure that DMZs are configured appropriately to prevent attacks that could exploit these such as zone pivoting.

Manufacturing Zone - Levels 0-3:

- Level 3 - Site Level: This level represents the highest industrial process level. Systems and applications at this level are responsible for managing site-wide industrial automation and control functions. This includes systems and functions such as plant historians, site-level operations management, control room workstations, file servers, and staging areas. Systems at this level are used to make changes to lower levels, such as patching and share data to the Enterprise Zone. While part of the operational process, most systems operating at this level are primarily based on standard equipment and operating systems such as Unix-based OSs or Microsoft Windows and, therefore, at a host level, have a low likelihood of being affected by tools and techniques employed during testing. However, if devices at this level are taken offline, cascading effects on lower levels need to be considered.

- Cell/Area Zone - Levels 0-2:
  - Level 2 - Area Control: This level contains systems and equipment responsible for the operation of an area within the industrial environment. This includes systems and equipment such as HMIs, alarms or alerting systems, and control room workstations. These systems communicate with ICSs in the basic control level (level 1) and share data to the site level (level 3). While most systems at this level are based on standard operating systems such as Unix-based OSs or Microsoft Windows, they also use industrial protocols to communicate with ICSs in level 1, which needs to be considered during scoping of adversary-centric security tests.
  - Level 1 - Basic Control: This level contains controllers that control and manipulate the operational process. Their primary function is to interface with process level (level 0) devices. This includes devices such as Programmable Logic Controllers or Distributed Control Systems. Most of these devices are based on proprietary operating systems, which are programmed and configured from upper-level workstations. These devices communicate to the level 0 devices they control and upper-level devices such as HMIs. This level has a high chance of causing disruption during an adversary-centric security test due to the direct impact on the operational process if any devices at this level are affected.
  - Level 0 - Process: This level contains a wide variety of sensors and actuators directly involved in the operational process. This can include devices with varying complexity, such as temperature gauges or a moving robot on an assembly line. Most of these use proprietary technologies to communicate with ICSs from level 1. Because of this, testing at this level is also considered high-risk.

Safety Zone:

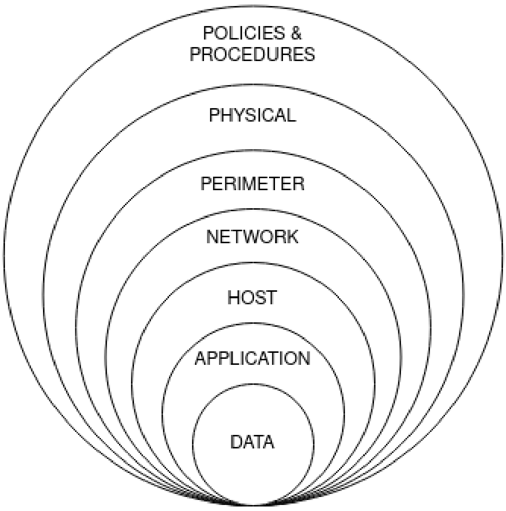

Figure 16: Defense in Depth Model

- Due to industrial environments having high safety requirements, safety equipment such as Safety Instrumented Systems (SIS) will reside within a zone separate from the Manufacturing Zone. Dedicated infrastructure is also implemented for their use to ensure their smooth operation. Disruption to endpoints within the Safety Zone during an adversary-centric security test is considered the highest risk due to their function of preventing loss of safety if disruption occurs in the Manufacturing Zone.

The second model used for developing the TiDICS methodology is the Defense in Depth (DiD) Model, illustrated in Figure 16. While the historical military strategy revolved around using weaker perimeter defence to allow the time to plan for a counter-attack, the cyber security strategy for DiD, conceived by the United States National Security Agency, involves parallel systems of physical, technical and administrative countermeasures to minimise the probability of a malicious actor gaining complete control of an environment [33]. Developed initially as a defensive strategy, the DiD model's layers can also be used as a testing methodology during adversary-centric security tests. These layers are as follows:

- Policies and Procedures: Cyber Threat Intelligence, Threat Modelling, Security Awareness Training, Security Governance, Risk Management, etc.

- Physical: Physical Access Control, CCTV, etc.

- Perimeter: Perimeter IDS/IPS/Firewall, DMZs, etc.

- Internal Network: Enterprise Remote Access, Content Filtering, Network Access Control, Data Loss Prevention, etc.

- Host: Patch Management, Endpoint Security Enforcement, Host IDS/IPS/Firewall, etc.

- Application: Database Monitoring, Dynamic/Static Application Testing, Application Firewall, etc.

- Data: Data Classification, Data Integrity Monitoring, Encryption, etc.

By combining these two models, The traditional testing methodology derived from the DiD model is now PERA Zone and Level dependent. For each layer of the DiD model, the safety and operational risks within each zone or level from the Purdue model also need to be considered during scoping of an adversary-centric security test. With this additional separation of zones and levels, scoping can be done based on identified safety and operational risks, allowing for further depth of testing of zones and levels with fewer risk factors.

By adding PERA zone and level requirements to the traditional testing methodology for the DiD model, scoping of adversary-centric security tests can be granularised into separate testing levels with varying degrees of risk to the operational process. Therefore, tools and techniques used for testing can be defined on a per-zone and per-level basis, allowing for extensive depth of testing while ensuring that risk is minimised for each of these.

6.2. Framework for Risk-Based Scoping of ICS/OT Adversary-Centric Security Tests

By applying the methodology for identifying and assessing safety and operational risk of adversary-centric security testing, described in Section 4 and the TiDICS model for defining testing of zones and levels with various risk factors, the following risk-based adversary-centric security testing framework is proposed. The core output of this framework provides a methodology for integrating safety and operational risk into the scoping adversary-centric security tests within ICS/OT environments; the extended framework (with example methodologies) is provided in Figure 18, and its process flow is provided in Figure 17. Figure 18 also distinguishes the different types of contributions that have been incorporated into the framework. These contributions have been classified into three different categories as follows:

- Blue: existing methodology applied to a novel context (e.g. applying (C)HAZOP to adversary-centric security testing)
- Green: existing methodology applied to established context (e.g. cut set probability for risk quantification)
- Red: novel methodology applied to novel context (e.g. data collection techniques for quantifying basic event risk)

A summary of the contributions that form part of the scoping framework, detailing how each contribution is used within the framework and the type of contribution, is provided in Table 3.

The overall framework is used sequentially as the output of previous phases is used as input for subsequent phases. The following subsections provide a description of these phases, their input requirements and their outputs.

6.2.1. Select TiDICS layers

Depending on the type of adversary-centric security tests and the budget of the organisation being tested, relevant PERA zones and levels can be selected to facilitate scoping of these. Once these zones and levels have been selected, subsequent DiD layers can be selected for identifying and quantifying safety and operational hazards for each of these. For example, The entirety of the Cell/Area Zone can be selected for scoping of a security test. Following this identification, only network and host DiD layers are selected for testing. This signifies that an assessment of safety and operational risks for the following zones and layers needs to be undertaken to scope the engagement: Cell/Area network; Area Supervisory Control Network and Host; Basic Control Network and Host; and Process Network and Host. Cell/Area Host testing is arbitrarily removed from scoping as the scoping for level 0 to 3 host testing implicitly results in that of the overall zone.

6.2.2. Identify Safety and Operational Hazards

Identifying Safety and Operational Hazards that can be caused due to active adversary-centric security testing for each of the selected TiDICS layers needs to be undertaken next. Several methodologies exist for identifying risk events and can be used at the framework user's discretion. An example of identifying these risk events using a (Control) Hazard and Operational Study ((C)HAZOP) is provided in Section 4.1 and demonstrates how risk events can be identified using a guideword methodology. This phase is primarily qualita-

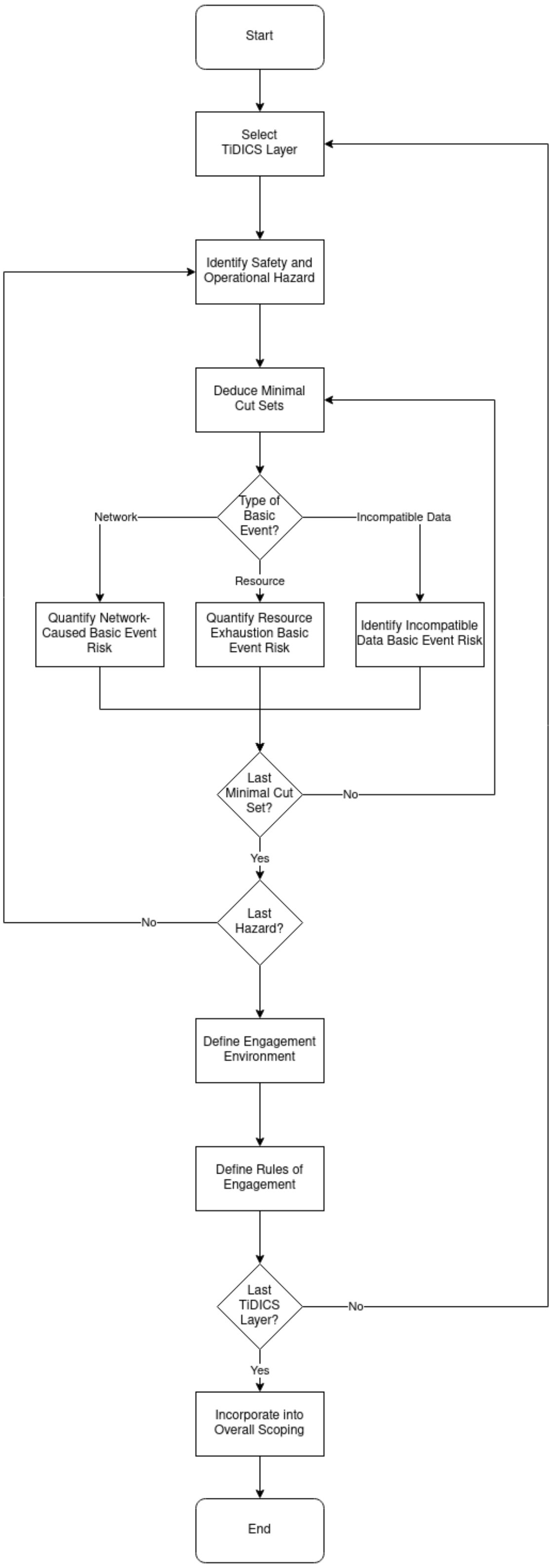
Start
Select TiDICS Layer
Identify Safety and Operational Hazard
Deduce Minimal Cut Sets
Type of Basic Event?
Network
Resource
Incompatible Data
Quantify Network-Caused Basic Event Risk
Quantify Resource Exhaustion Basic Event Risk
Identify Incompatible Data Basic Event Risk
Last Minimal Cut Set?
No
Yes
Last Hazard?
No
Define Engagement Environment
Define Rules of Engagement
Last TiDICS Layer?
No
Yes
Incorporate into Overall Scoping
End

Figure 17: Scoping Framework Process Flow

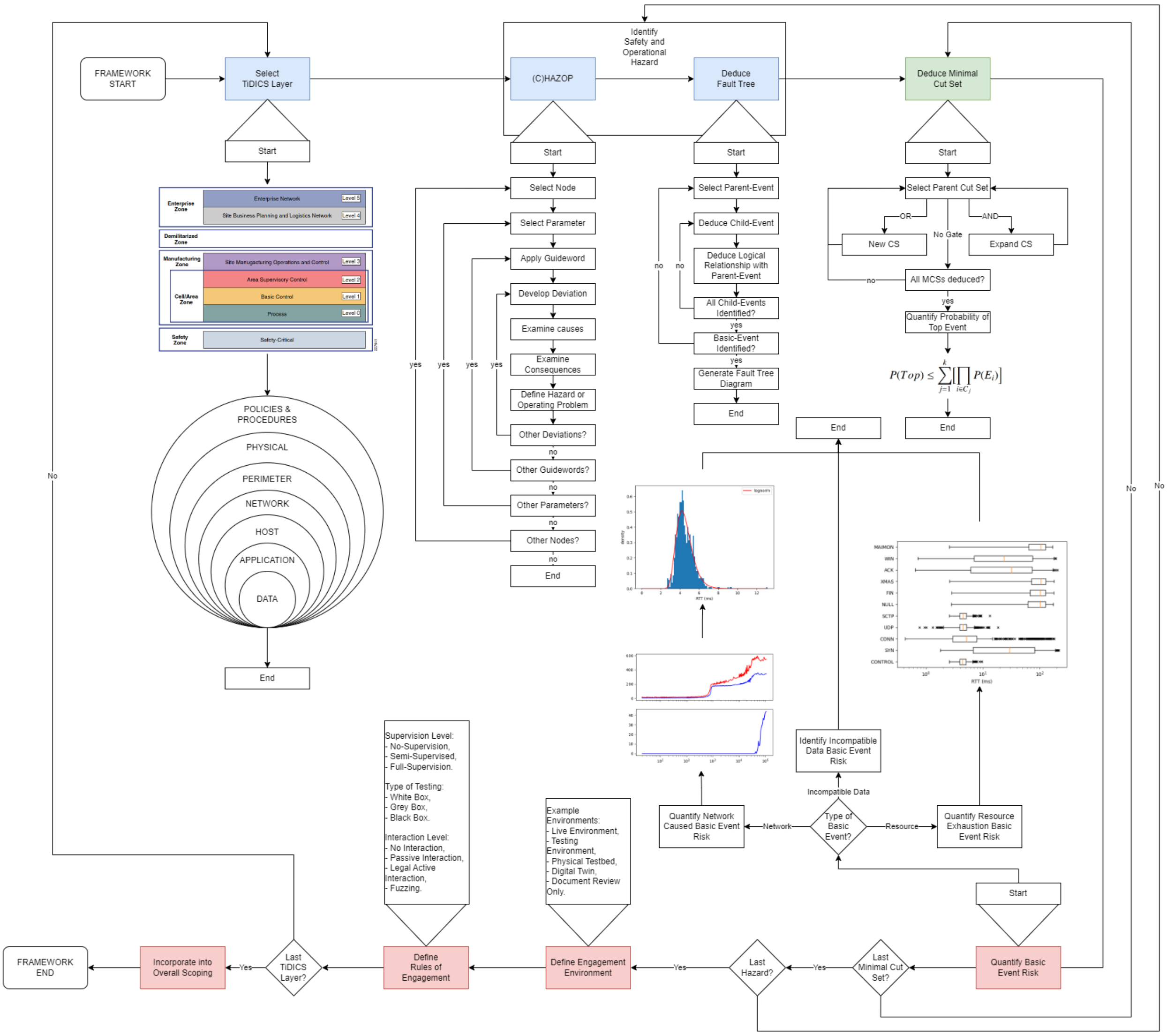

Figure 18: Extended Framework for Safety-Risk-Based Scoping of Adversary-Centric Security Tests

Contribution Name	Contribution Summary	Methodology Type	Context Type
TiDICS Model	Model for selecting testing zones with different safety/operational risk requirements	Existing	Novel
(C)HAZOP	Identifying safety/operational risks of security testing	Existing	Novel
FTA	Decomposing safety/operational risks of security testing	Existing	Novel
MCS probability	Quantifying safety/operational risks of security testing	Existing	Existing
Basic Event Data	Methods for collecting data for risk quantification	Novel	Novel
Engagement Environment	Defining the engagement environment	Novel	Novel
Rules of Engagement	Defining the rules of engagement	Novel	Novel
Incorporation	Incorporating output in overall scoping methodology	Novel	Novel

Table 3: Summary of Contributions Forming the Scoping Framework

tive and relies on existing documentation, such as P&I diagrams, network diagrams, configuration documents, and other relevant documents, for deducing risk events.

6.2.3. Decompose Safety and Operational Hazards

Once safety and operational hazards have been identified, these can be further decomposed into a combination of basic events that, if happen simultaneously, lead to a top event, or major hazard, occurring. Again, while the specific methodology for doing this is subject to the user's discretion, the framework provides an example Fault Tree Analysis methodology, discussed in Section 4.2. Conducting an FTA allows framework users to generate minimal cut sets of basic events, which can be represented through boolean algebra and therefore used to quantify risk precisely.

6.2.4. Quantify Network-Caused Basic Event Risk

Basic events that contribute to a reduction or loss in availability caused by latency increase or packet loss are categorised as Network-Caused Basic Events. The probability of these can be determined through prior testing or estimated by expert opinion. By determining the maximum allowed network throughput for specific endpoints or networks, the probability of tools or techniques affecting availability at these throughputs can be determined by considering network jitter. This probability of disruption to the operational process can be determined and used to determine appropriate security testing tools.

6.2.5. Quantify Resource Related Basic Event Risk

Basic events that contribute to a reduction or loss in availability caused by resource exhaustion are categorised as Resource Exhaustion Basic Events. The probability of these can be determined by testing tools and techniques planned to be used during the security test and determining their effect on the target. If tools have multiple options for performing a similar task, these can be compared to determine the options that produce the most negligible overhead and are the safest for use.

6.2.6. Identify Incompatible Data Basic Event Risk

Basic events that contribute to a loss of integrity or a total loss in availability caused by incompatible or anomalous data being sent to a target and unable to be processed or understood are incompatible Data Basic Events. Identifying risk mitigation techniques for specific tools or techniques that cause these basic events can be done through logical experimentation (i.e. changing tool options or improving exception handling on targets). However, if no solutions can be identified within an appropriate time frame, these tools and techniques must be documented, and their use during the security test should be prohibited to prevent disruption to the operational process.

6.2.7. Define Engagement Environment

Based on the quantified risks of selected adversary-centric security testing tools and techniques, the environment for deploying these can be selected. Such environments that can be used for testing include but are not limited to: live environment (high inherent risk and high accuracy of testing results), testing environment (medium inherent risk and accuracy of testing results), physical testbed (low inherent risk and medium-low accuracy of testing results), digital twin (low inherent risk and low accuracy of testing results), and document review (no inherent risk and very-low accuracy of testing results). For example, if the risks quantified from previous steps are low, testing can be conducted either in a live environment or a test environment to ensure maximum depth of testing. However, if the risks quantified from previous steps are high, testing can be conducted in lower-risk environments such as testbeds to prevent disruption to the operational process.

6.2.8. Define Rules of Engagement

Based on the selected engagement environment, rules of engagement must be defined and subsequently enforced during the entirety of the security test. Several types of rules of engagement can be defined, such as supervision level (no-supervision, semi-supervision or continuous supervision), the type of testing (white box, grey box, or black box), the type of interaction with endpoints (no-interaction, passive interaction or active interaction) and more if required. These rules of engagement are categorised and defined within general scoping methodologies for adversary-centric security tests using comprehensive risk treatment principals, found in ISO 31000 [34], for example. These risk treatment options are not always mutually exclusive or appropriate in certain scenarios and can include risk avoidance, risk acceptance, risk removal, reducing likelihood, impact modification, or risk sharing. As described in ISO 31000, the use of risk treatment should take into consideration all party's obligations, voluntary commitments, views from stakeholders, objectives, risk criteria and available resources.

6.2.9. Incorporate into Overall Scoping Methodology

Finally, these results can be documented for use in the overall scoping of an adversary-centric security test to understand the safety and operation risks that are present. Incorporating the outputs of the framework in the overall scoping of the test ensures maximal depth-of-testing and minimal disruption to the operational process.

7. Evaluation

We have introduced a framework to aid stakeholders in scoping adversary-centric security tests within ICS/OT by quantifying safety and operational risk. The framework provides a step-by-step methodology for identifying safety and operational hazards, decomposing these hazards into risks, quantifying these risks, and defining scoping constraints for integration into the overall scoping methodology of these engagements. While the framework is supported by commonly adopted methodologies and data collected from the Lancaster University ICS testbed, these only provide a proof of concept of how the framework could be adopted and implemented in practice. Therefore, in this section an evaluative study of the framework to determine its potential for implementation in practice is presented.

While current literature and practice acknowledge that safety and operational risk are a concern when conducting adversary-centric security tests in ICS/OT environments, these do not address how to identify and understand these risks to reduce the impact of these engagements on safety and the operational process. The risk-based safety scoping framework was created to address this gap to enable safe adversary-centric security tests within ICS/OT environments.

7.1. Evaluation Design

There are risks related to the usage of the proposed framework, these are primarily due to the possible lack of safety and operational risk analysis maturity from target users. To this end, an evaluation of the implementation of the framework in practice is required to assess its validity, accuracy, and applicability based on existing processes.

To address this, we opted for semi-structured interviews with key stakeholders for the data collection part of the evaluation: ICS/OT cyber security consultants, engineers, and penetration testers. These roles have a direct bearing on the framework's implementation. The nature of semi-structured interviews strikes a balance between flexibility and precision. Unlike structured interviews, they offer the adaptability needed to capture the nuances of the participants' experiences, while being more structured than unstructured interviews. This duality ensures that while discussions can be tailored to individual respondents, they remain anchored to core evaluation themes [35]. Consequently, this approach paves the way for template analysis of the gathered data, as posited by King [36]. Our primary aim was to discern the prevailing implementations of safety risk analysis techniques and gauge the applicability of the scoping framework in these contexts.

While powerful in drawing broad generalizations, quantitative methodologies might fall short in capturing the richness and depth required to assess the practical application of our framework. While quantitative methods offer measurable, broad-scale data, they can lack the depth and subjective insights crucial for our study. As [37] and [38] suggest, qualitative research provides a richer, more detailed view, which is especially valuable in studies like ours where expert opinion helps shape understanding. Further to this, there is an understandable reluctance to apply our framework into an operational environment without the required risk evaluations and other safety-related precautions. Semi-structured interviews allow us to augment our quantitative testbed evaluation used to design the framework to obtain practical insights into its deployment that go beyond our predicted outcomes [37]. Our study depends on understanding experts' qualitative feedback, which offers insights into the accuracy, reliability, validity, and applicability of the framework and these interviews allow us to delve deeply into the topic and achieve a thorough understanding of the answers provided [39]. Specifically, it helps explain the complexities of implementing the framework in real-world scenarios and brings forth potential challenges or constraints that might hinder its broad adoption. Quantitative approaches tend not to provide such an intrinsic viewpoint into the human aspect.

Our qualitative approach, focusing on semi-structured interviews with industry experts, was chosen to evaluate the framework holistically, ensuring that its attributes align with real-world safety and operational risk dynamics.

7.1.1. Methodology

To properly conduct our semi-structured interviews, we needed to develop a clear methodology to be used for the interviews. This was focused on who the participants would be and which questions we needed to ask to obtain the data required for the analysis. This approach also ensured better consistency between the interviews and in the subsequent analysis. Due to the critical nature of the context in which the scoping framework is intended for, a high level of confidence is required before it can be tested and evaluated in real industrial environments. To this end, a qualitative approach was selected in which interviewing stakeholders across the topic area of ICS/OT adversary-centric security testing was applied. For this, participants were provided with the framework and an example application of its use, using data collected from the Lancaster University testbed running the same scenario as presented in Section 4. This approach provides two benefits: firstly, participants were able to provide their opinion on the framework, its phases, as well as any observations on potential limitations that could hinder its use in practice; and secondly, using the data collected from the aforementioned scenario, participants could provide insight on the framework's accuracy, reliability, validity, and applicability for implementation in practice. These evaluation criteria have been defined as follows:

- Accuracy: How close is the output of the framework to the correct and accepted outcome?
- Reliability: How does repeated use of the framework affect its outcome when provided the same input?
- Validity: How appropriate is the framework in addressing its objective?
- Applicability: How much does the outcome of the framework change when used in different contexts?

Interviews were selected as an appropriate method that enables each participant to discuss their personal experience concerning the risks of adversary-centric security testing in ICS/OT environments and how the framework can be used to address these concerns [40]. For this, a semi-structured approach was also adopted as it provides adequate flexibility with a core question set while allowing the option to include improvised follow-up questions for further exploration of topics of interest [35]. Furthermore, the threats to validity concerning the reliability of the collected data using this method have previously been addressed [4].

Participant Selection. The aim of selecting an appropriate participant sample is to understand the topic area from all relevant perspectives. To achieve this, a broad approach was applied to target participants. This resulted in a diverse collection of role profiles. More specifically, roles that would engage in the scoping or implementation of adversary-centric security tests within ICS/OT environments across multiple backgrounds and with varying levels of responsibilities. This sampling approach provides multiple perspectives, building an accurate perspective on the scoping framework's validity for implementation in practice.

To summarise, five participants were selected holding the following roles:

- ICS/OT Security Researcher
- ICS/OT Cyber Security Engineer
- Health and Safety Manager
- Operations and Finance Chief Information Security Officer
- Filling and Packing Operations and Cyber Security Manager

The levels of experience varied amongst participants within each of the defined roles ranging from five to forty-six years; the majority of which, however, had been working with industrial systems for over ten years.

While having a sample size of five for this evaluation may seem insufficient for the accurate evaluation of the framework, the main purpose of this evaluation is to identify limitations with its implementation in practice. Nielsen et al. [41] provide a mathematical model for finding usability problems that can be used to plan the amount of evaluation required to achieve the desired level of thoroughness. This work demonstrates that by conducting a qualitative study with five participants, 85% of the issues in a proposed work, such as the scoping framework, in this case, will be identified. Ensuring that a 31% chance exists that each participant will identify an issue if it exists, is also defined as a requirement. Given the role profiles and the levels of experience of the selected participants for the study, this requirement is met as each person has sufficient expertise to identify any issues with the framework. Therefore, having a sample size of five for evaluating the framework is appropriate for identifying at least 85% of the possible limitations concerning the implementation of the framework in practice.

7.1.2. Interview Protocol/Guide

Each interview was broken down into the following seven stages, providing a logical structure to the interview protocol/guide:

- Preface
- Establishing Demographics
- Framework Familiarisation
- Framework Evaluation
- Scenario Familiarisation
- Framework Evaluation (With Application Scenario)
- Conclusion

The core focus of these interviews was to present participants with the risk-based safety scoping framework for adversary-centric security testing on ICS/OT. More specifically, how key stakeholders would approach risk quantification for the scoping of these engagements using the framework. The questions aligned to these interview stages are aided through the inclusion of probes and definitions.

Establishing Demographics Phase. The following question-set was applied to the demographics phase:

- Please can you tell us your job title and provide a brief overview of your core roles and responsibilities?
- How many years experience do you have working in this role?
- At a very high level, please can you explain what you understand the term adversary-centric security test to mean?
- Have you ever been involved in an adversary-centric security test that was performed for an ICS/OT environment?
- At a very high level, what do you believe to be the greatest challenges of conducting adversary-centric security tests for ICS/OT environments?

Framework Evaluation Phase. The following question-set was applied during the Framework Evaluation phase once participants had been presented with the risk-based scoping framework:

- What is your opinion on using the TiDICS model for separation of testing zones and layers based on safety and operational risks?
- Do you agree that the types of risk that adversary-centric security testing presents to safety and the operational process are comprehensively considered within the framework?
- What challenges could affect the collection or quality of data for risk quantification?
- From the framework's overview, do you think the output of the framework could be used in the overall scoping of an adversary-centric security test?
- From the framework's overview, do you believe that the output of the framework is accurate enough to ensure a full understanding of the safety and operational risks from adversary-centric security testing on ICS/OT so that depth of testing can be maximised while minimising risk to the operational process?
- From the framework's overview, do you believe that the framework can be applied in all ICS/OT environments where safety and operational risks are a concern?

Framework Evaluation (With Application Scenario) Phase. The following question-set was applied during the Framework Evaluation (With Application Scenario) phase once participants had been presented with an example scenario for applying the framework:

- Does your opinion of the framework's accuracy, reliability, validity, and applicability change when presented with an example application of its use?
- What is your opinion on the use of (C)HAZOP for identifying hazards that could occur during an adversary-centric security test on ICS/OT?

- What is your opinion on the use of FTA to decompose hazards into smaller basic events for use in risk quantification?
- What is your opinion on the methodologies used for quantifying the risk of basic events?
- Overall, would you use this framework as part of the overall scoping methodology for an adversary-centric security test on ICS/OT?

Conclusion Phase. The following question-set was applied during the conclusion phase:

- Would you like to add anything which may be relevant?

7.1.3. Analysis

Template analysis was selected to analyse the interviews, as it is a highly flexible method for analysing qualitative data [42]. This approach is considered a middle ground between the relatively rigid content analysis approach in which analytical codes are all predefined [43] and the opposite approach of grounded theory in which all analytical codes must be derived from the data [44]. This methodology can create an initial code set based on the interview protocol/guide, which aligns with the core areas of interest for evaluating the framework. Furthermore, template analysis allows for creating additional code sets to analyse discussion points previously not considered.

7.2. Results

Building on the methodology and design outlined in previous sections, we could analyse the data from our semi-structured interviews. Our focus here was to consolidate and interpret the qualitative data obtained from these interviews, emphasizing the real-world applicability, challenges, and effectiveness of the framework. We structured this analysis around key themes identified within the interviews and confirmed by means of an thematic analysis.

7.2.1. Challenges of Adversary-Centric Security Testing on ICS/OT

Before evaluating the framework, it is important to understand what participants understand to be the current challenges of conducting adversary-centric security tests within ICS/OT environments. Participants' opinions aligned closely with the findings in section 6. Due to the nature of ICS/OT environments, any engagement that can potentially affect the operational process can lead to a loss of business continuity or even safety.

'When you are operating an OT environment, everything needs to be run as efficiently and effectively as possible and any slight deviations from that can not only put human safety at risk, but it can also completely throw off the operational process and in turn cost a lot of money.'

The safety and operational risks are amplified due to the design philosophies of ICS. Generally, these are designed to favour environmental resilience and operational longevity instead of performance. These limited resources could affect availability in the context of an adversary-centric security test if using tools that are resource heavy. The timing for conducting adversary-centric security tests within ICS/OT environments was also discussed. Due to the high up-time requirements of these environments, maintenance periods for live environments are uncommon, which can occur every five to ten years in some environments. Because of this timeline, a decision has to be made on whether engagements can and should be conducted outside maintenance periods or should be restricted to these.

Due to ICS/OT cyber security, in general, being an interdisciplinary field, several different skill sets are required for successfully conducting adversary-centric security tests. Therefore, ensuring that these different actors are present was identified as crucial by participants to ensure all required considerations during the planning and execution of these engagements. This requirement is equally applicable to the test providers, which also require ICS/OT knowledge in order to successfully understand how to provide the correct services for these environments. Because of the wide variety of protocols and product vendors used for ICS/OT, difficulties also arise when conducting adversary-centric security testing within environments that incorporate a broad range of protocols and device types. This variety adds additional complexity to these engagements due to different protocols and devices often being incompatible with each other. For example, if Modbus and S7COMM were used within the same network, this could require different tools for conducting tests which leverage these protocols, such as SimaticScan [13] for S7COMM.

'Because you have different types of devices, if you have a unique way to test these with one tool, that would be helpful for the OT team and allow for comparable results between devices.'

7.2.2. Selection of Testing Zones Using the TiDICS Model

Despite the Purdue Model and DiD Model not being initially intended for scoping adversary-centric security tests, participants agreed that these could be used for defining testing areas for safety and operational risk analysis. While used as a reference architecture, the Purdue Model separates zones and levels based on hierarchical function. Because of this, the different zones and levels that can be selected, using the TiDICS model, will also contain different risks, enabling distinct risk analysis processes for each of these.

'I think it provides you the capability of at least defining which devices you're going to test. And definitely the further down [the Purdue Model levels] the more things are critical.'

As discussed in Section 6, while the Purdue Model is commonly adopted for designing ICS/OT networks, other reference architectures are also used in practice, such as the one recommended by IEC 62443 [45]. For this, the framework can be modified to select testing zones based on these reference architectures, instead of or in addition to the TiDICS model, allowing flexibility for environments that use different network architectures than the one described in the Purdue Model.

'I would definitely look at how IEC 62443 handles zones and conduits because I think it would make the framework even more accessible for environments that are already following 62443.'

Similarly, the DiD Model, while used for implementing different defensive controls and policies on a per-layer basis, can also be used to further granularise the selection of areas to test. The risk for each of these layers can then be analysed for use in subsequent phases of the framework.

A few participants noted the dependency between certain layers of the DiD model depending on the design philosophies of the networks or devices being tested. For example, when testing the application layer of a target device, this dependency might imply that testing at a host level overall is also required due to the architecture of the target device(s). The TiDICS Model could facilitate risk identification and analysis for different environments with a similar configuration. This would allow framework users to streamline specific steps of the framework for multiple environments, reducing the cost of the scoping process for adversary-centric security tests.

'It allows us to deep dive into the risk of a specific system and replicate that somewhere else with similar configurations. It allows us to do a safety risk assessment that's reproducible in similar environments.'

7.2.3. Framework Users

While the framework was initially intended to be used by both the test providers and the environment asset owners (including IT and OT engineers), a third party was identified that would need to be involved in the scoping process to maximise its efficiency: the product/solutions vendor. This is because vendors provide certain environments as part of a black-box solution.

Because of this, certain organisations' engineering teams may understand how their environments function but would need to gain the required knowledge concerning the inner workings of specific devices for in-depth risk analysis. This lack of required information would also include crucial documents used for risk identification, such as network diagrams, which are essential for comprehensively considering the risk that conducting an adversary-centric security test could present to these environments.

'For example, the network diagrams - we don't have that. It's the vendor that has that. They could give us that information if required but often it's them that understands fully how the PLCs and networks are configured.'

7.2.4. Safety and Operational Hazard Identification

As part of the initial phase of the framework for safety and operational risk identification, difficulties could arise for organisations without proper asset management maturity. This lack of asset management would impact the quality of hazard identification and could lead to significant hazards not being identified.

'if we actually know what we have in the network. [...] You need to know what you've got in order to be able to analyse it.'

Identification of hazards would also need to include the product/solutions vendor to fully understand the risk when conducting an adversary-centric security test. While only a few participants were aware of (C)HAZOP prior to the demonstration of the framework, its ease of use and high-level operation allow it to be used within several contexts, including adversary-centric security testing. Applying guidewords to parameters to identify potential deviations allows framework users to comprehensively consider all risks that could be present during such an engagement.

'It's interesting in the sense that we take each parameter and apply a guideword to identify hazards. [...] It's more precice than say "if the HMI doesn't

work anymore, what happens". Because CHAZOP provides such a high level methodology, it can be applied within a lot of different contexts, which is why it's good to use. '

Participants appreciated that the framework offered flexibility on the methodologies for identifying hazards, enabling the use of methodologies already used for traditional safety risk assessment in the context of adversary-centric security testing. Using hazard identification methodologies that are well established and already used by the framework users also facilitates this phase of the framework since a new methodology would not need to be learnt from the beginning. For organisations that do not have an established methodology for identifying hazards, providing (C)HAZOP as an example methodologies offers framework users a starting point in the event that more guidance is required for this phase of the framework.

7.2.5. Risk Initiator Deduction

To enable risk quantification in subsequent phases of the framework, identified hazards need to be broken down into smaller quantifiable risks. While not all participants were familiar with Fault Tree Analysis, most participants were familiar with methodologies that use tree diagrams, such as attack trees or probability trees.

FTA was deemed adequate for decomposing hazards into smaller risks as it provides a logic tree that describes relations and dependencies between different risks. Being environment-agnostic, it is also widely applicable to several contexts, including adversary-centric security testing. Because FTA uses logic tree diagrams, these can be directly translated to probability tree diagrams. Subsequently, these can be used to calculate the probability of identified hazards occurring based on smaller quantifiable risks.

'FTA is a powerful tool to take into accounts all the devices, steps, causes and consequences. So for me, it's adaptable and the power of that is that it's not linked to any specific environment. You can use it for areas like filling and packing, manufacturing, flows, etc.'

Similar to how (C)HAZOP is provided as an example methodology for identifying risks, decomposing hazards can be done using other methodologies if they achieve the same result of being able to quantify risks. However, FTA was accepted as a suitable recommended methodology and should be used if there are no established methodologies for decomposing risk. One concern that was raised was the scalability of using these methodologies for adversary-centric security tests with large scopes. While the scenario provided was simple in concept, some industrial environments can be very large and complex, which could lead to very large and complex fault trees being generated. In this case, the framework does acknowledge scalability issues and suggests mitigating these by using fault tree generation software to automatically create FTDs.

7.2.6. Collecting Data

In order to efficiently quantify the risks of conducting adversary-centric security tests within ICS/OT environments, the appropriate data needs to be collected. However, the skill gap between different fields of expertise could lead to vital information being missed for risk quantification. This highlights the importance of including all relevant parties in the scoping process so that the collected data comprehensively covers all the risks present during these engagements.

'It's very difficult to find somebody who's an expert in OT, risk and cyber security at the same time. So for me one of the main issues is generally the knowledge of the people involved in the scoping process. Who needs to be involved so that it [the collection of data] can be done to an in depth extent?'

While passively collecting data in live environments is possible, some environments might be provided as black-box solutions and require the product/solutions vendor to provide appropriate data or install solutions for passively collecting data, such as maintenance sensors. In some instances, however, the vendor may be unwilling to implement these. In order to safely collect data for quantification of risk through active methods, doing so through a staging environment or testbed was recommended in order to prevent potential disruption within a live environment.

The quality of collected data from an isolated environment could impact the preciseness of the quantified risks. While this is true, it was generally accepted that if the offline environment (such as a testbed) closely mirrors the live environment, the margin of error in risk quantification would be negligible. However, this margin of error reduces dramatically for risks with high impacts. If a PLC crashes or hangs during an offline test, for example, it is likely to react the same within the live environment. Identifying the appropriate data to collect is also important. With modern technology, however, the amount and quality of the collected data could cause delays during this phase of the framework. Therefore, understanding which data is vital for risk quantification is crucial to prevent unnecessary complications.

‘What data would actually be useful for evaluating risk? How do we collect this? How would it be used to perform an analysis? If it’s an old system, do we actually have the capability to collect data for this? [...] Since ICS/OT, when implemented, is designed to stay operational for years, that could make things difficult.’

7.2.7. Methods for Risk Quantification

When discussing example methodologies for quantifying the risk of basic events, the results from these would allow testers to determine precisely how tools and techniques could affect safety or the operational process. In doing so, additional information could be identified, such as the effect of network-based attacks on an environment.

The depth of analysis provided by the example methodologies lead a few participants to question their current process for scoping adversary-centric security tests within ICS/OT environments. By conducting a more in-depth analysis with quantifiable results, risk can be precisely calculated and used in the scoping process, improving the depth of testing while minimising risk to safety and the operational process. Mapping the initial objectives of the adversary-centric security test to the methods used for risk quantification was also identified as vital. In doing so, associated risks can be minimised to prevent disruption to business continuity risk or vendor maintenance risk, and testing quality can be improved.

‘This allows us to use the correct tools for the given scope but we would need to associate them with the initial objectives that were defined. For me, there’s two things that we need to reduce: the risk associated with business continuity and vendor maintenance risk. If the vendor says: “your pen-test has changed everything based on our initial configuration and that means we can’t ensure proper maintenance of the network” then that’s very bad.’

While some scalability issues were identified for environments containing a wide variety of protocols and device types, participants generally agreed that the methodologies provided in the framework were a good starting point for quantifying safety and operational risk and should be used for the scoping of adversary-centric security tests on ICS/OT environments.

7.2.8. Framework Outputs

Once safety and operational risks have been quantified, these can be used to define scoping constraints for use in the overall scoping methodology of an adversary-centric security test on ICS/OT environments. Such constraints include defining the engagement environment, for example. While testing on a live environment is possible, providing that the risk is sufficiently low, a few participants stated that some stakeholders may still be too risk-averse to allow any engagement despite this low risk.

‘I’m not sure how far people want to do tests on the live environment just yet. Although it might be more of a possibility in the future given more open-minded people and more modern technology.’

On the contrary, some participants stated that they would only be able to conduct tests in a live environment due to the lack of a staging area or testbed. Again, this emphasises the need to quantify safety and operational risks so that scoping constraints can enable safe adversary-centric security tests within live environments. For this reason, another constraint identified as essential for scoping adversary-centric security tests is when to conduct the engagement. While specific environments may have long timeframes between maintenance periods, these can be used to conduct security tests with lower risk to the operational process than outside of these periods.

Since risk is quantified in previous phases of the framework, precise scoping constraints can be defined, such as limiting network throughput on specific tools. This allows framework users to make accurate decisions that can be re-evaluated post-engagement for future security tests.

7.2.9. Framework Discussion

Overall, participants were mostly receptive to the framework and its use in the risk-based safety scoping of adversary-centric security tests on ICS/OT environments. However, a few limitations were identified with this. Firstly, the maturity of framework users may impact its effectiveness. Because of the critical nature of ICS/OT environments, stakeholders may be unwilling to conduct security tests within their environments even if the quantified risk is low. However, with modern technologies and methodologies for quantifying these risks, the framework presents an opportunity for safely conducting adversary-centric security tests within ICS/OT environments.

‘It’s not a limitation of the framework itself, it’s a limitation on the maturity of the users of the framework. I probably would say that most people

would be a little too cautious than actually they should be. The more people use the framework and the more knowledgeable they become about this sort of thing, then the more people might be willing to take risks of doing these tests in these environments.'

As discussed in previous sections, scalability was also identified as a potential framework limitation. The resources and time required for effective scoping of adversary-centric security tests may be too costly for large and complex environments. However, the framework provides users with a starting point for quantifying safety and operational risks to enable security tests in these engagements, especially for organisations that require such tests as part of compliance requirements.

While the framework is intended initially to aid in scoping adversary-centric security tests for ICS/OT environments, it also serves as a tool to increase an organisation's asset management maturity. This framework can, therefore, be used in other areas, such as general safety risk management or improving operational asset resilience.

'It's not just enabling a penetration test, it's discussing and finding out information about the assets within an environment. Not only would clients be provided with a penetration test at the end of this, they'd also know more about their assets and it would enrich their asset register.'

Despite the framework requiring technical competence for its operation, participants appreciated that it was presented in a way that could be understood at all skill levels. This allows the framework to include technical and non-technical users in scoping adversary-centric security tests within ICS/OT environments.

'It's simple and I don't mean that as in it's easy. You've managed to make a very complex process become simple. It seems to be adequate to explain the scoping to both engineers and the security operations centre for example - and even senior (management) people. It would probably put their mind at rest.'

While the framework provides flexibility in the methodologies used for risk identification and quantification, it offers recommended methodologies, including (C)HAZOP and FTA, that can be used following the provided guidance. The framework's efficiency can also be improved the more it is utilised through lessons learnt. For example, assessing whether the defined objectives prior to scoping the adversary-centric security test have been met can be used to improve upon future engagements and the quality of risk scoping. Finally, most participants shared that, having been presented with the framework, they would insist that such a process must be used prior to conducting adversary-centric security tests within their ICS/OT environments.

'I would actually insist that something like this was done, if it could be done. I think that it could be a requirement during the call of offers from pen-test providers to ask that they can ensure that the tools they are going to use aren't going to affect business continuity.'

8. Discussion

Our framework introduces a method for integrating the quantification of safety and operational risks into the process of planning adversary-focused security tests within ICS/OT settings. It employs a flexible strategy, enabling users to adapt various risk assessment methods to suit their particular requirements. These methods provide guidance for stakeholders to use their current risk assessment practices in the planning of adversary-centric security tests, effectively quantifying safety and operational risks.

During the development of the framework, we were mindful of certain limitations that might hinder its effectiveness. A notable concern is the framework's reliance on the TiDICS model, which combines aspects of the Purdue and Defense in Depth Models. Although the Purdue Model is widely accepted as a benchmark for bolstering environmental security via zone separation, not every organisational network architecture aligns with it. This variance adds complexity to the process of selecting zones and levels for risk quantification. We advise that organisations with inadequate security-focused network architectures or underdeveloped asset management practices rectify these issues before embarking on adversary-centric security testing. For environments that are modelled after the Purdue Model or similar structures, like those outlined in the IEC 62443 [45], the TiDICS model can be modified to accommodate these variations.

Implementing (C)HAZOP and FTA for risk assessment comes with its own set of challenges. The qualitative essence of (C)HAZOP necessitates its integration with FTA to enable effective risk quantification. Organisations need to have a certain degree of maturity in asset management and established safety risk assessment practices for these methods to produce accurate results.

Both (C)HAZOP and FTA have their strengths in certain aspects of risk analysis, but they might not be comprehensive enough to fully cover common cause effects, which often involve multiple failures or conditions leading to a single consequence. These effects require an analysis that can encompass complex interdependencies and simultaneous failures. For this reason, (C)HAZOP and FTA are provided as example methodologies only within the framework rather than requirements for the operation of the framework. Alternatives such as Systems Theoretic Process Analysis (STPA) [46], which considers safety risks as issues of control rather than failure, offer more precision in control-intensive environments and take into consideration common cause-effects. Additionally, Event Tree Analysis (ETA) can be used in conjunction with FTA. This forward-looking approach can help in understanding the different outcomes that can arise from a single initiating event, including common cause failures [47, 48]. Experiment specification [49] can also be used alongside the proposed framework to provide further depth in assessing the security capabilities of target environments.

We realised that deploying this framework in real-world environments would require substantial trust from operators. Consequently, we chose a qualitative evaluation method, conducting semi-structured interviews with industry experts. This approach enabled us to collect diverse industry viewpoints on the framework, evaluating its practicality and efficacy.

Participant feedback suggested that the stages of the framework were realistic and tackled the challenges of performing adversary-centric security tests in ICS/OT settings. The adaptability of the TiDICS model in accommodating various reference architectures, such as IEC 62443, was especially valued. A critical aspect that was initially overlooked in the current version of the framework was the role of product/solution vendors in the planning process. Including them could significantly enhance the identification and quantification of risks, thereby notably improving the output quality of the framework. The framework's recommendation to employ existing organisational methods for hazard identification and risk reduction was well received, as was the incorporation of (C)HAZOP and FTA as suggested methodologies. These methods assist in establishing clear scoping constraints, facilitating comprehensive adversary-centric security testing while reducing risks to safety and operations.

Despite the favourable feedback, some concerns were raised about the practicality of the framework due to the maturity level of its potential users. While this is more a limitation of the users than of the framework itself, we have countered this by providing detailed guidance for each stage of the framework. Scalability in large, complex environments remains a challenge; nevertheless, the framework offers a valuable foundation for integrating safety and operational risk quantification in such scenarios.

In summary, the evaluations affirmed the framework's precision, dependability, validity, and suitability. The interviews confirmed that the framework significantly improves the planning of adversary-centric security testing, potentially enriching an organisation's asset management process beyond its original intent.

9. Conclusion and Future Work

With this paper we propose a scoping framework for adversary-centric security testing. It incorporates identified safety and operational risks into the overall scoping process to minimise the risk of disrupting the operational process. This has been done with a focus on maximising the depth of testing where possible. Our frameworks also includes the proposal of a hybrid testing methodology, named the Testing in Depth for ICS (TiDICS) model. We achieved this through the combination of the Purdue Model [32] and the Defense in Depth Model [33].

To provide a validation of the proposed framework semi-structured interviews were conducted with industry stakeholders, including penetration testers and Operational Technology asset owners. The framework was identified as a significant step forward in improving the safety and security of industrial environments and was deemed to be applicable in practice, based on the opinion of experts. Although some limitations were identified, such as the scalability of the framework, these can be mitigated through further research or the integration of additional tools.

The interviews conducted in Section 7 identified that further research could be conducted regarding the framework's scalability for large and complex environments. However, several ways exist to address this and ensure that the cost of resources does not outweigh the framework's benefits. For example, through repeated use of the framework, data from previous engagements can be used to enhance the subsequent quantification of risks. This can reduce the need to repeat certain phases, such as risk identification or data collection, thereby saving resources. Another approach to reducing the cost of resources required for the operation of the framework is to leverage automated risk analysis and security testing, a field that has seen considerable research [50, 51, 52]. By integrating automated methods into specific framework phases, such as risk identification or data collection, its efficiency can be improved to reduce the required resources for its operation. With these strategies in place, the framework can be enhanced to be a valuable tool for scoping adversary-centric security tests within more environments, including large and complex ones.

Finally, the evaluation also highlighted the potential benefits of conducting further quantitative evaluation of the framework to reinforce its effectiveness. This approach would build on the conducted quantitative and qualitative evaluation and provide additional confidence for its use in real environments. Initial steps have already been taken towards further quantitative evaluation through data collection within our ICS testbed. Conducting further evaluations in real industrial environments would provide tangible evidence of the framework's applicability in practice and further reinforce its value to users.

References

[1] T. Miller, A. Staves, S. Maesschalck, M. Sturdee, B. Green, Looking back to look forward: Lessons learnt from cyber-attacks on industrial control systems, International Journal of Critical Infrastructure Protection 35 (2021).

[2] Centre for the Protection of National Infrastructure, Critical National Infrastructure, https://www.cpni.gov.uk/critical-national-infrastructure-0, 2021. Last Accessed: 2022-06-16.

[3] C. Vallance, Ukraine war: Major internet provider suffers cyber-attack, https://www.bbc.com/news/60854881, 2022. Last Accessed: 2022-06-16.

[4] A. Staves, T. Anderson, H. Balderstone, B. Green, A. Gouglidis, D. Hutchison, A Cyber Incident Response and Recovery Framework to Support Operators of Industrial Control Systems, International Journal of Critical Infrastructure Protection (2022). URL: \url{https://www.sciencedirect.com/science/article/pii/S187454822100086X}. doi:https://doi.org/10.1016/j.ijcip.2021.100505.

[5] B. Green, D. Prince, J. Busby, D. Hutchison, "How Long is a Piece of String": Defining Key Phases And Observed Challenges within ICS Risk Assessment, Association for Computing Machinery, New York, NY, USA, 2017. URL: https://doi.org/10.1145/3140241.3140251. doi:10.1145/3140241.3140251.

[6] A. Staves, A. Gouglidis, D. Hutchison, An Analysis of Adversary-Centric Security Testing within Information and Operational Technology Environments, Digital Threats: Research and Practice (2022).

[7] C. Sherry, Advantages and Disadvantages of Active vs. Passive Scanning in IT and OT Environments, 2020. URL: https://bit.ly/3trOgLy, last Accessed: 14-06-2021.

[8] Siemens, SIMATIC S7-1200, https://new.siemens.com/global/en/products/automation/systems/industrial/plc/s7-1200.html, ???? Last Accessed: 2022-06-16.

[9] Allen Bradley, Controllogix controllers, revision 16, https://literature.rockwellautomation.com/idc/groups/literature/documents/rn/1756-rn016_-en-e.pdf, ???? Last Accessed: 2022-06-16.

[10] B. Green, R. Derbyshire, W. Knowles, J. Boorman, P. Ciholas, D. Prince, D. Hutchison, ICS Testbed Tetris: Practical Building Blocks Towards a Cyber Security Resource, in: 13th USENIX Workshop on Cyber Security Experimentation and Test (CSET 20), 2020.

[11] J. Gardiner, B. Craggs, B. Green, A. Rashid, Oops i did it again: Further adventures in the land of ics security testbeds, CPS-SPC'19, Association for Computing Machinery, New York, NY, USA, 2019, p. 75–86. URL: https://doi.org/10.1145/3338499.3357355. doi:10.1145/3338499.3357355.

[12] M. Dietz, M. Vielberth, G. Pernul, Integrating digital twin security simulations in the security operations center, in: Proceedings of the 15th International Conference on Availability, Reliability and Security, ARES '20, Association for Computing Machinery, New York, NY, USA, 2020. URL: https://doi.org/10.1145/3407023.3407039. doi:10.1145/3407023.3407039.

[13] R. Antrobus, S. Fey, B. Green, A. Rashid, SimaticScan: Towards A Specialised Vulnerability Scanner for Industrial Control Systems, in: 4th International Symposium for ICS & SCADA Cyber Security Research 2016, 2016.

[14] D. Efanov, Plcscan, https://code.google.com/archive/p/plcscan/, 2012. Last Accessed: 2022-06-27.

[15] E. Samanis, J. Gardiner, A. Rashid, A taxonomy for contrasting industrial control systems asset discovery tools, 2022. URL: https://arxiv.org/abs/2202.01604. doi:10.48550/ARXIV.2202.01604.

[16] B. Green, R. Derbyshire, M. Krotofil, W. Knowles, D. Prince, N. Suri, PCaaD: Towards Automated Determination and Exploitation of Industrial Systems, Computers & Security 110 (2021) 102424.

[17] S. Maesschalck, A. Staves, R. Derbyshire, B. Green, D. Hutchison, Walking under the ladder logic: PLC-VBS: a PLC control logic vulnerability scanning tool, Computers & Security 127 (2023) 103116.

[18] I. Schieferdecker, J. Grossmann, M. Schneider, Model-based security testing, arXiv preprint arXiv:1202.6118 (2012).

[19] S. Hollerer, T. Sauter, W. Kastner, Risk assessments considering safety, security, and their interdependencies in ot environments, in: Proceedings of the 17th International Conference on Availability, Reliability and Security, 2022, pp. 1–8.

[20] S. Hollerer, W. Kastner, T. Sauter, Towards a threat modeling approach addressing security and safety in ot environments, in: 2021 17th IEEE International Conference on Factory Communication Systems (WFCS), IEEE, 2021, pp. 37–40.

[21] T. Kletz, I. of Chemical Engineers (Great Britain), Hazop & Hazan: Notes on the Identification and Assessment of Hazards, Hazard workshop modules, Institution of Chemical Engineers, 1986. URL: https://books.google.fr/books?id=RrGUQgAACAAJ.

[22] J. Love, Hazard Analysis - Process Automation Handbook: A Guide to Theory and Practice, Springer London, 2007. URL: https://doi.org/10.1007/978-1-84628-282-9_54. doi:10.1007/978-1-84628-282-9_54.

[23] J. Dunjó, V. Fthenakis, J. A. Vílchez, J. Arnaldos, Hazard and Operability (HAZOP) Analysis. A Literature Review, Journal of Hazardous Materials 173 (2010) 19–32. URL: https://www.sciencedirect.com/science/article/pii/S0304389409013727. doi:https://doi.org/10.1016/j.jhazmat.2009.08.076.

[24] S. Whitty, T. Foord, Is HAZOP Worth all the Effort it Takes?, Technical Report, Institution of Chemical Engineers, 2009.

[25] American National Standards Institute / International Soci-

ety of Automation, ANSI/ISA-5.1-2009 - Instrumentation Symbols and Identification, http://integrated.cc/cse/Instrumentation_Symbols_and_Identification.pdf, 2009.

[26] International Organization for Standardization, ISO 14617-6:2002 - Graphical symbols for diagrams Part 6: Measurement and control functions, https://www.iso.org/standard/22654.html, 2002.

[27] International Electrotechnical Commission, IEC 61025:2006 - Fault Tree Analysis, https://webstore.iec.ch/publication/4311, 2006.

[28] A. Staves, ICS/OT testing data sets and scripts, https://github.com/Warschak/ICS-OT-testing-data-sets-and-scripts, 2022.

[29] M. Karakas, Determination of Network Delay Distribution over the Internet, Master's thesis, The Middle East Technical University, 2003.

[30] M. Mozhaiev, N. Kuchuk, M. Usatenko, The method of jitter determining in the telecommunication network of a computer system on a special software platform, Innovative Technologies and Scientific Solutions for Industries (2019) 134–140. doi:10.30837/2522-9818.2019.10.134.

[31] E. Daniel, C. White, K. Teague, An interarrival delay jitter model using multistructure network delay characteristics for packet networks, in: The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003, volume 2, 2003, pp. 1738–1742 Vol.2. doi:10.1109/ACSSC.2003.1292282.

[32] P. Didier, F. Macias, J. Harstad, R. Antholine, S. A. Johnston, S. Piyevsky, D. Zaniewski, S. Zuponcic, M. Schillace, G. Wilcox, Converged Plantwide Ethernet (CPwE) Design and Implementation Guide, Rockwell Automation 9 (2011) 564.

[33] National Security Agency, Defense in Depth: A practical strategy for achieving Information Assurance in today's highly networked environments, https://web.archive.org/web/20121002051613/https://www.nsa.gov/ia/_files/support/defenseindepth.pdf, 2012. Last Accessed: 12/07/2022.

[34] International Organization for Standardization, ISO 31000:2018 - Risk management — Guidelines, https://www.iso.org/standard/65694.html, 2018.

[35] H. Arksey, P. T. Knight, Interviewing for social scientists: An introductory resource with examples, Sage, London, 1999.

[36] N. King, Template analysis, in: G. Symon, C. Cassell (Eds.), Qualitative methods and analysis in organizational research: A practical guide, Sage Publications Ltd., 1998, pp. 118–134.

[37] N. K. Denzin, Y. S. Lincoln, Introduction: The discipline and practice of qualitative research. (2008).

[38] J. W. Creswell, C. N. Poth, Qualitative inquiry and research design: Choosing among five approaches, Sage publications, 2016.

[39] M. C. Harrell, M. Bradley, Data collection methods: Semi-structured interviews and focus groups (2009).

[40] M. Q. Patton, Qualitative evaluation and research methods, SAGE Publications, inc, 1990.

[41] J. Nielsen, T. K. Landauer, A mathematical model of the finding of usability problems, in: Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems, Association for Computing Machinery, 1993, p. 206–213. URL: https://doi.org/10.1145/169059.169166.

[42] C. Cassell, G. Symon, Qualitative methods in organizational research: A practical guide, The Qualitative Research Interview 17 (1994).

[43] R. P. Weber, Basic content analysis, volume 49, Sage, 1990.

[44] B. Glaser, A. Strauss, Grounded theory: The discovery of grounded theory, Sociology the journal of the British sociological association 12 (1967) 27–49.

[45] IEC, IEC 62443, 2019.

[46] I. Friedberg, K. McLaughlin, P. Smith, D. Laverty, S. Sezer, Stpa-safesec: Safety and security analysis for cyber-physical systems, Journal of Information Security and Applications 34 (2017) 183–196. URL: https://www.sciencedirect.com/science/article/pii/S2214212616300850. doi:https://doi.org/10.1016/j.jisa.2016.05.008.

[47] J. Ignac-Nowicka, Application of the fta and eta method for gas hazard identification for the performance of safety systems in the industrial department, Management Systems in Production Engineering 26 (2018). doi:10.2478/mspe-2018-0003.

[48] R. Ferdous, F. Khan, R. Sadiq, P. Amyotte, B. Veitch, Fault and event tree analyses for process systems risk analysis: Uncertainty handling formulations, Risk analysis : an official publication of the Society for Risk Analysis 31 (2011) 86–107. doi:10.1111/j.1539-6924.2010.01475.x.

[49] P. Smith, E. Piatkowska, E. Widl, F. P. Andrén, T. I. Strasser, Towards a systematic approach for smart grid hazard analysis and experiment specification, in: 2020 IEEE 18th International Conference on Industrial Informatics (INDIN), volume 1, IEEE, 2020, pp. 333–339.

[50] T. Hussain, R. Eschbach, Automated fault tree generation and risk-based testing of networked automation systems, in: 2010 IEEE 15th Conference on Emerging Technologies & Factory Automation (ETFA 2010), 2010, pp. 1–8. doi:10.1109/ETFA.2010.5641309.

[51] J. I. Single, J. Schmidt, J. Denecke, State of research on the automation of HAZOP studies, Journal of Loss Prevention in the Process Industries 62 (2019) 103952. URL: https://www.sciencedirect.com/science/article/pii/S0950423019302323. doi:https://doi.org/10.1016/j.jlp.2019.103952.

[52] A. Applebaum, D. Miller, B. Strom, C. Korban, R. Wolf, Intelligent, Automated Red Team Emulation, Technical Report, The MITRE Corporation, 2016.

Appendix A. Best-Fit Results of Network Jitter Distribution using `distfit`

ranking	distr	score	LLE	loc	scale	arg
0	genextreme	0.04456	NaN	4.124234	0.715471	(0.019880619231587585,)
1	lognorm	0.048644	NaN	1.341275	3.059706	(0.2670700729681621,)
2	gamma	0.055465	NaN	2.094081	0.30557	(7.918591272636394,)
3	beta	0.055789	NaN	2.127674	1787830.197773	(7.6293048670266215, 5714159.193619767)
4	t	0.111718	NaN	4.43861	0.710655	(6.171637809106455,)
5	dweibull	0.117561	NaN	4.46801	0.712976	(1.2178260986783505,)
6	norm	0.135643	NaN	4.513763	0.899412	()
7	loggamma	0.146162	NaN	-354.035654	45.900016	(2469.654144604755,)
8	expon	0.982031	NaN	2.653	1.8607632	()
9	uniform	1.326598	NaN	2.653	10.4656	()
10	pareto	1.384878	NaN	0.001607	2.651393	(1.9463309032892653,)

Table A.4: Best Fit Results using `distfit` for PLC Network Distribution at 400 packets per second

ranking	distr	score	LLE	loc	scale	arg
0	lognorm	2.496344	NaN	0.686484	0.208651	(0.519992978790028,)
1	genextreme	2.646364	NaN	0.861475	0.084229	(-0.16375062691567044,)
2	beta	2.824626	NaN	0.714093	1592192285724.914062	(2.8199955923958058, 21167823283185.176)
3	dweibull	5.24911	NaN	0.886558	0.098125	(1.0069936788017444,)
4	t	7.237047	NaN	0.907803	0.101204	(5.579147878141951,)
5	norm	9.109383	NaN	0.926964	0.168039	()
6	loggamma	9.893515	NaN	-51.624706	7.135857	(1579.2097368296409,)
7	expon	16.320509	NaN	0.719	0.207964	()
8	pareto	17.57609	NaN	-1085520.353355	1085521.072355	(5560848.1067601815,)
9	gamma	33.248012	NaN	0.719	0.47821	(0.1306685429804617,)
10	uniform	39.455348	NaN	0.719	3.175	()

Table A.5: Best Fit Results using `distfit` for HMI Network Distribution at 400 packets per second