Digital Twin-Enhanced Incident Response for Cyber-Physical Systems

ABSTRACT

Cyber-physical systems underpin many of our society's critical infrastructures. Ensuring their cyber security is important and complex. A major activity in this regard is cyber security incident response, whose primary goal is to detect and mitigate cyber-attacks in order to ensure the continuity and resilience of services. For cyber-physical systems this is particularly challenging because it requires insights both from the cyber and physical (process) domains and the engagement of stakeholders that are not strictly concerned with cyber security. A technology that is receiving a lot of attention are digital twins – virtual representations of real-world (cyber-physical) systems. They can be used to support tasks such as estimating the state of a system and exploring the consequences of interventional activities (e.g., upgrades).

In this paper, we examine the use of digital twins to support cyber security. Specifically, our novel contribution is to provide a comprehensive analysis of the types of activities and how different modalities of digital twin use can be applied to the phases of cyber security incident response. Building on this analysis, we propose a structured approach to enhancing cyber security playbooks for cyber-physical systems incident response with digital twins. Playbooks are an essential component of incident response, ensuring that multi-disciplinary teams are effective in responding to cyber security incidents; therefore, improvements in their execution can result in increased resilience. To illustrate our approach, we present its use for a playbook that is concerned with mitigating a cyber-attack to critical industrial equipment.

CCS CONCEPTS

• **Computer systems organization → Embedded and cyber-physical systems**; **Heterogeneous (hybrid) systems**; • **Computing methodologies →** *Simulation tools.*

KEYWORDS

Digital twins, incident response, cyber-physical systems, playbooks, security operations center

ACM Reference Format:
. 2023. Digital Twin-Enhanced Incident Response for Cyber-Physical Systems. In *ARES '23: International Conference on Availability, Reliability and Security, August 29 – September 1, Benevento, Italy*. ACM, New York, NY, USA, 10 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Industrial Control Systems (ICS) are becoming increasingly digitised due to the adoption of Industry 4.0 concepts and the shift towards so-called "smart manufacturing". This has led to the increased prevalence of Cyber-Physical Systems (CPS) – potentially large, heterogeneous systems consisting of physical machinery, Information Technology (IT) and Operational Technology (OT). The importance of securing CPS must not be understated as they are a core element in many critical infrastructures. The Stuxnet attack, Trisis/Triconex malware, and the 2015 attack on Ukraine's energy grid are just some of the incidents that have demonstrated not only the vulnerability of CPSs to cyber-attack, but the incredibly serious and potentially life-threatening consequence of attacks against them [21, 23, 28].

Incident Response (IR) in CPSs is a particularly challenging aspect of ensuring safe and secure operation. As a result of digitisation, parsing, organising, and prioritising security events – many of which can be false positives – becomes increasingly difficult due to the increase in attack surface area. The adoption of cloud computing results in data being transported and stored off-site, often via third-party vendors, while embedded technologies and new industrial automation protocols often increase the number of potentially exploitable services running on plant equipment. Moreover, securing a CPS is also an extensive undertaking, as they consist of many different forms of physical machinery, processes, and networked computer systems. Stringent availability requirements for CPSs means that downtime is kept to a minimum due to safety concerns, as well as financial or reputational loss. This reduces the amount of time available for in-depth Digital Forensics and Incident Response (DFIR) activities. This challenge is further complicated by the varying levels of security maturity of IT and OT systems, resulting in a wide range of expertise needed to operate, monitor, and secure a CPS.

These challenges make it difficult to understand the state of the CPS – particularly its digital components. An under-utilised technology spawned from Industry 4.0 concepts that can aid practitioners throughout the IR life cycle is the *Digital Twin* (DT). The DT concept – modelling or representing real-world entities in virtual space – has been thoroughly explored for predictive maintenance, efficiency analysis, and product design, by applying several different models or modelling techniques, and using real-time and historical data from real systems. DTs also have the benefit of being able to explore a range of realistic system states and scenarios that may not be possible to explore in the real-world system due to financial and temporal costs, or environmental risk. One area of DT application that is still being explored is the use of DTs for cyber security applications. A key gap in this area of research is the integration of DTs in the IR life cycle, including existing SOC workflows.

ARES '23, August 29 – September 1, 2023, Benevento, Italy
© 2023 Association for Computing Machinery.
ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

In this paper, we propose the use of DTs for cyber security activities throughout the stages of the IR Life Cycle [22]. This proposal is presented in Section 4. Previous research is categorised according to the IR phase and *mode of operation* (i.e. type) of the presented DT. In Section 5, we present a novel approach for integrating DTs into IR playbooks by examining existing procedures to identify common workflow steps, such as *scope adjustment* and *containment*. These workflow steps can be supported with a DT using a set of well-defined query types – aligned with Pearl's hierarchy of causal inference [25] – that correspond to identified DT modes of operation. The goal is to produce an enhanced IR playbook that incorporates DTs in order to make their execution more effective and less error prone, for example. We illustrate this approach in Section 6 with an example playbook for a cyber-attack against a Programmable Logic Controller (PLC). In what follows, we present important related work and essential preliminaries regarding the modes of DT operation for cyber security and incident response.

2 RELATED WORK

Eckhart and Ekelhart define a DT in the context of CPS security [14] as being a "virtual replica of a system that accompanies its physical counterpart during phases of its life cycle, that consumes real-time and historical data if required, and has sufficient fidelity to allow the implementation of the desired security measure". With this definition in mind, we present a review of relevant literature in this area.

Vielberth et al. propose the use of DTs as an input to the creation of a cyber range for training Security Operations Center (SOC) staff [27]. The authors describe the work of Bécue et al. [4] as using a DT *as a* cyber range, in contrast to Vielberth et al. who propose the use of a DT as a valuable input *in the creation of* a cyber range, and not as a cyber range itself. The authors examine the three modes of operation of a digital twin — data analysis and optimisation, simulation, and replication [9] — but through the lens of how beneficial it is to the creation of a cyber range for training SOC staff. As a result, the authors place less value on data driven approaches as they cannot produce the level of immersion required for training. Vielberth again, alongside Dietz and Pernul (see [9, 10]), published preceding work that sought to demonstrate how simulation of security events in a DT can provide a SOC with helpful insights and "support the enhancement of SIEM systems" [11]. They exhibit a framework using Business Process Model and Notation (BPMN) which focuses on the DT, SOC, and the SIEM tool of an organisation, and primarily focus on simulating cyber-attacks. In contrast, our proposal uses existing incident response practices and workflows to demonstrate how a wide range of DTs can aid incident response for a range of stakeholders.

Further integration of DTs into IR strategies is explored in [16] with DTs being proposed as part of Security Orchestration, Automation, and Response (SOAR) tool to update device firmware or remove devices from a network automatically based on a pre-planned playbook response. The authors implement a digital twin using the Eclipse Ditto IoT DT framework that provides a cloud-based view of real-world system data. Their approach focuses solely on the eradication and recovery IR phases as it aims to address automation and response (i.e. SOAR).

Dietz et al. propose a replication-based DT for digital forensics [8]. Data is taken from real-world system and fed to a DT. System states are replicated at every step to ensure that the virtual system mirrors the real-world system. The virtual system can be used for digital forensics before the real-world forensics takes place, saving time (and money) during the forensics process.

Allison et al. propose the use of DTs in a methodology that addresses the problem of data scarcity for cyber security machine learning models in CPSs [1]. The methodology also covers the commissioning of machine learning models — trained on DT-generated data — on edge-based devices. A DT of a NPP is used to generate a range of allowable and abnormal operating conditions that are used to train and test an anomaly detection model, respectively.

Eckhart et al. exhibit a framework based on DTs that aims to improve "cyber defence capabilities" of CPS operators [15]. The framework provides visualisations and replays recorded system states for the operators to better understand the "cyber situation". The authors also emphasise the important role that visualisations play in providing useful security-related information. The authors extend Eckhart and Ekelhart's *CPS Twinning* framework for providing an environment for a DT based on the CPS's specification [12].

The reviewed works have mostly focused on single a use case or phase within the IR life cycle. While there has been work on specific IR activities, the full life cycle has not yet been fully considered. Furthermore, the reviewed research does not adequately integrate into existing cyber security practices. Our paper addresses this gap by examining the use of the DT throughout the full IR life cycle, analysing existing IR processes, and demonstrating a method for integrating DTs into IR playbooks.

3 PRELIMINARIES

In this section, we summarize key concepts that form the basis of our contribution – they relate to cyber security incident response for CPSs and the different modes of operation of DTs.

3.1 Digital Twin Modes of Operation

Dietz and Pernul identify three modes of operation (MO) of a *security-focused* digital twin – data analysis and optimisation, simulation, and replication [9]. This is a well-cited organisation of DTs for cyber security that we have adopted for our work. A brief summary of each MO is provided.

Data Analysis and Optimisation. DTs operating for data analysis and optimisation (or simply "Data Analysis") make use of technologies and techniques including machine learning, statistical analysis, time series analysis, regression, forecasting techniques, etc. Historical data can be used to baseline or train models to understand normal behaviour (or, in cases where labelled data is available, also abnormal behaviour). These analysis methods are used to detect abnormal scenarios as well as extrapolate trends and predict certain events in the future.

Simulation. Simulation-based DTs are based on models derived from specifications and measurements from real-world objects and systems. They are initiated with data from the real-world system, as well as user-specified parameters. They are executed to observe

Table 1: Literature referenced in Section 2 presenting DTs relating to cyber security; organised by mode of operation (MO) and the most significant contributions pertaining to each stage of the SANS Incident Response Life Cycle.

	Preparation	Identification	Containment	Eradication	Recovery	Lessons Learned
Data Analysis.	[1]	[1]	[16]	[16]	[16]	[1]
Simulation	[1, 9, 11, 27]	[1]	[2, 4, 11]	[2, 4]	[2, 4, 11]	[2, 9, 11, 27]
Replication	[13, 15]	[8, 15]	[8]	[8]	[8]	[15]

how systems and their environment interact with one another, enabling users to observe phenomena such as cascading effects of faults and cyber-attacks, or conditions that are unable to be tested in the real-world due to the risk of equipment failure or human safety concerns.

Replication. Replication-based DTs are also designed based on specification data; however, their goal is to mimic the real-world system as closely as possible in order to provide a digital copy for the user to explore. They take input from the real-world system and attempt to reproduce real-world conditions, relying more on their design and specification than historical data. Deviation from the specification is easily highlighted with replication-based digital twins. In [9], an example of the advantage of this form of DT is given in the context of the Stuxnet attack [28].

3.2 Incident Response

IR is approached by way of analysing and collating evidence from potentially affected systems. Based on this evidence, actions are taken to limit the number of affected hosts or systems, and address the root cause of a problem. IR is divided into cyclical phases. These phases are described in the the SANS Incident Response Cycle [22], which consists of six phases; expanding upon the four phases of the NIST Incident Response Life Cycle [5]. We have chosen to align this paper with the SANS approach, as *Containment*, *Eradication*, and *Recovery*, are sufficiently different in their aims and approach to warrant their own individual phases.

Table 1 shows how the literature referenced in Section 2 aligns with the SANS Incident Response Cycle. The six phases of this cycle include: preparation, identification, containment, eradication, recovery and lessons learned, and can be summarized, as follows:

Preparation includes activities that improve the speed or timeliness of identification of threats and intrusions and increase the ability to handle threats at any moment.

Identification includes identifying (and confirming) intrusions and other security events. This stage will involve the collection of evidence from different sources such as host logs, firewalls, intrusion detection/prevention systems (IDS/IPS), etc.

Containment addresses the source of the intrusion to ensure it does not worsen or spread, and includes short-term containment, forensic imaging, system backup, and long-term containment.

Eradication includes removing the different traces of intrusion, including malicious files or devices, and restoring the system to an uncompromised state. It also addresses the source of the problem, whether a compromised user account, attacker back door,or needing to rebuild a new system to avoid the consequences of the incident in the future.

Recovery includes replacing or restoring the components that are required to return the system back to a trusted or clean state, returning as close to a pre-intrusion state as possible so that the system can operate as intended. Restored systems should be monitored to ensure the effects of the intrusion have been remedied and that the system has in fact been cleaned and patched to prevent the intrusion occurring again.

Lessons Learned includes documenting and analysing the incident to gain insights on the incident response process to refine or redesign processes, edit documentation, and adjust methods of identification when necessary.

Many phases of IR are executed by following *playbooks* of known or common scenarios in order to increase the efficiency of incident handling, thus limiting potential damage (reputational, financial, physical) [7]. Playbooks contain workflows in the form of flowcharts or step-by-step instructions that provide technical details on how to remedy a situation. Workflows aim to reduce the amount of subjective human reasoning (e.g., gut feeling responses, curiosities, etc.) and limit the human user to answer a series of almost purely objective “yes/no" questions, thereby reducing human error in high-pressure situations, such as responding to cyber-attacks against critical business functions.

CPSs form the basis of critical infrastructure, therefore there are many stakeholders invested in their safe and secure operation. This list includes operations personnel, SOC personnel, maintenance staff (often in the form of third-party contractors), Original Equipment Manufacturers (OEMs), the National Computer Emergency Response Team (CERT) (or equivalent), and national authorities (police, national security services, etc.). These stakeholders may be engaged at different points of a playbook's execution, and their involvement can be critical in determining the root cause and remedy of an incident.

4 DIGITAL TWIN APPLICATIONS FOR INCIDENT RESPONSE

In order to understand how DTs can be integrated into IR workflows, we need to understand what phases can benefit from the different digital twin MOs (recall Section 3.1). After analysing current uses of DTs for cyber security in both literature (see Table 1) and industry, we produced a non-exhaustive list of DT applications that can be offered throughout the IR life cycle by DTs for each MO – data analysis and optimisation, simulation, and replication. This is shown in Table 2.

Throughout the preparation and identification IR phases, data-driven DTs can be used to analyse system interdependencies or

Table 2: Phases of Incident Response (IR) and some examples of applications (use cases) that can be offered by Digital Twins (DTs) of different Modes of Operation (MO).

IR Phase	Data Analysis & Optimisation	Simulation	Replication
Preparation	- Training data driven models on historical system data or DT data. - Analysing system interdependencies from DT data and prioritising systems for protection in the event of fault or cyber-attack.	- Practicing IR in DT. - Prioritising systems and services for protection based on outcome of cyber-attack simulation. - Prioritising systems for safe system shutdown based on simulated faults. - Validating hazard analysis via simulation. - Training system operators during fault/cyber-attack scenarios.	- Using replicated digital systems to identify areas where evidence can be found for digital forensics. - System snapshots can be used for training users of real-world system.
Identification	- Real-time, data-driven analytics and anomaly detection.	- Deviations from simulated behaviour can indicate signs of cyber-attack/anomaly. - Root cause analysis.	- DT derived from specification can be used to highlight compromise/anomaly.
Containment	- Automated shutdown of infected device. - Automated updating/editing of firewall rules to prevent connection to/from a compromised machine. - Automated switching to redundant, isolated controller.	- Shutting down vulnerable/exploited service on a server can be tested in a DT. - Testing containment related firewall rules. - Testing updated SIEM-ingested IOCs, Yara rules, etc. - Testing safe system shutdown or system isolation in DT.	- Forensic imaging of replicated system before performing DFIR activities on the real system.
Eradication	- Automatic triggering of anti-malware software scans on suspected compromised systems.	- Updating and patching vulnerable systems. - Restoring OS from an original disk image from the vendor.	- Scanning replicated systems with anti-malware programs.
Recovery	- Data-driven monitoring of systems that have been restored.	- Testing patches for vulnerable systems and services in DT before patching real systems. - Integration testing restored systems in a DT before restoring real system. - Monitoring restored systems via simulation.	- Monitoring replication-based DTs of restored systems.
Lessons Learned	- Auto-documenting incidents using data from data-driven DT.	- Simulation results can be used in post-incident documentation for justification of decisions made by SOC analyst or operations personnel.	- Snapshots of digital systems can be used for post-incident analysis.

provide real-time anomaly detection or causal analysis. However, it is less clear how these types of DTs can be applied during the containment, eradication, and recovery phases, as these are more practical, hands-on phases that focus on interfering with the operation of the system. Simulation provides many IR phases with either a tool for side-by-side comparison with the real-world system or a tool for projecting its future states. It can also provide a test bed for any activity that involves intervening with the system, such as containment, eradication, and recovery. Meanwhile, replicated systems provide a baseline against which the real-world system can be compared. Furthermore, replicated systems – similar to simulations – allow the user to test interventions in a safe environment, including digital forensics. Replication-based twins that follow the real-world system strictly in real time and at the lowest levels can provide another monitoring point for the system. Moreover, the ability to snapshot the system can be useful for post-incident analysis (lessons learned) and training personnel (preparation).

5 DIGITAL TWIN ENHANCED PLAYBOOKS

Building on our analysis of the use of DTs for IR, in this section, we focus on how they can be applied to enhance the execution of IR playbooks. Initially, an analysis of emergency procedures is discussed. Based on this, we propose a categorization of playbook activities and then describe a process that can be used to determine how DTs can be integrated into playbooks.

5.1 Emergency Procedures Analysis

To integrate DTs into playbooks for IR, we are required to know what form the playbooks take and how the playbooks function in reality. A pre-cursor to modern SOC IR Playbooks include Abnormal/Emergency Operating Procedures (AOPs/EOPs), used to handle physical operations in critical infrastructure, including Nuclear Power Plants (NPPs). We analysed AOP/EOP design documents and examples from the International Atomic Energy Agency to better understand how the operators of critical infrastructure approach incident handling [18, 19].

To gain the SOC's perspective, we analysed example IR playbooks from Microsoft [24] and the United States Cybersecurity and Infrastructure Security Agency (CISA) [7], and playbook design specifications [6, 17]. These documents revealed common features that are used for IR in both the cyber and physical operations domains.

All playbooks contain an *initiating condition*, which is the reason for the playbook being executed, such as a SIEM alert or abnormal operating conditions. Detailed, step-by-step workflows within the playbook usually contains a series of common steps, including:

(1) Obtaining information from diverse, redundant sources, to assess the state of the system.
(2) An evaluation of system state via a series of *yes/no* questions.
(3) An adjustment of scope based on the evidence gathered.
(4) A pivotal decision point when an incident is confirmed or dismissed.
(5) A short-term containment strategy in the event of a confirmed incident.
(6) An assessment (or reassessment) of the severity of the incident.
(7) An in-depth analysis of affected systems.
(8) The eradication or removal of the root cause of the incident.
(9) Exit conditions or methods of incident escalation that conclude the workflow.

Throughout IR workflows cyber security, many stages have dedicated *documentation steps* for the analyst to pause and update SIEM/SOAR tools or ticketing systems with the new information or a new assessment of the incident. Secondary paths are offered for when access to primary tools or instrumentation is obstructed due to analyst permissions, system ownership, or system function. For example, a junior SOC analyst will not have access to a PLC controlling a subsystem in the primary loop of an NPP, due to the critical function of the PLC and the risk of creating a safety-related incident. Furthermore, permissions to make changes to devices may be limited to operations personnel or a third-party vendor.

Playbooks may contain extra information, such as references to relevant legal and regulatory requirements, along with descriptions of the mechanisms for involving appropriate stakeholders, such as contact information for the relevant personnel within national authorities.

This form of organised, proceduralised incident management enables clearer post-incident analysis, as the responses are predefined [17]; however, as Savioja et al. note, procedures intended to be objectively followed can be interpreted differently depending on the person or team implementing the procedure [26].

5.2 Playbook Activity Categorization

Derived from the analysis of IR playbooks and the activities in Table 2, we have identified six main areas of support provided by a DT during IR. These six areas of support – shown in Table 3 – fall into two main assistance categories: *Management and Oversight* and *Digital Forensics and Incident Response (DFIR) Assistance.* Within Management and Oversight there are two areas of support, namely *Communication* and *Documentation.* Within DFIR Assistance there are four areas of support, namely *Status Query*, *Associative Query*, *Interventional Query*, and *Counterfactual Query.* The associative, interventional, and counterfactual concepts are based on Pearl's three-level hierarchy of causal inference [25]. This section explains these concepts in the context of a DT.

5.2.1 Management and Oversight.

Documentation refers to the use of the DT platform for updating the playbook, online documents, incident reports, etc., with data and information automatically gained from the DT. For example, a replication-based DT built from a detailed system specification can automatically update incident reports by populating them with information on infected hosts. Not only does this provide information in easily digestible formats, supporting situational awareness, but documentation and report automation gives the human user more time for investigating the source of an anomaly. This feature of a DT could also integrate into – or exists as an extension of – already existing reporting mechanisms, such as case reporting in SIEM solutions.

Communication refers to the sending of messages, system data, system snapshots, simulation results, etc., between CPS stakeholders during an incident. This can be an automated or manual process. For example, operations personnel may initiate a series of faster-than-real-time simulations to forecast future system states during an incident. Reports generated by the DT can then be shared with other stakeholders, such as regulatory authorities or law enforcement.

5.2.2 DFIR Assistance.

A *Status Query* involves using a DT model to investigate the system state. This could include the use of replication-based DTs to investigate the system's intended design against the current system state. Anomalies could be found in the differences between the DT and the real-world system.

An *Associative Query* utilises data to provide statistical relationships. Such associations can be inferred by standard conditional probabilities and conditional expectation. Current machine learning methods are used for answering these questions. For example, if we observe a water leak in one subprocess, then a drop in pressure elsewhere is more likely or expected. This is Judea Pearl's first level in the hierarchy of causal inference.

An *Interventional Query* includes “what-if" questions and ranks higher in the hierarchy as it involves not just observed data, but

Table 3: The main areas of support provided by a DT in the execution of a playbook.

Assistance Category	**Management & Oversight**		**DFIR Assistance**			
Area of Support	Documentation	Communication	Status Query	Associative Query	Interventional Query	Counterfactual Query
Purpose	Supporting stakeholder engagement and streamlining IR reporting processes.		Reasoning about or investigating the current system state.		Forecasting system states and understanding effects of potential interventions.	

also changing parameters. For example, *what if* we opened the valves more to compensate for the water leak? Would we maintain pressure or run out of water? This form of query can make use of cyber ranges, replication-based DTs, high-fidelity simulations, etc. to model the system with new parameters.

A *Counterfactual Query* includes retrospective reasoning. If we have a model that can answer counterfactual queries, we can also answer questions about interventions and observations; therefore, it is at the top of the hierarchy. This form of query features more during post-incident analysis (*Lessons Learned* phase), but can also form part of operator training (*Preparation* phase).

5.3 Digital Twin Incident Response Playbook Integration

How DTs can be integrated into IR playbooks and workflows depends on the existing IR processes. If playbooks are yet to be designed or implemented in the organisation, there is an opportunity to consider the use of DTs during the design process and parallelise activities and order activities based on the availability of DT results and forecasts. Conversely – and the focus of this section – if there are mature, proven playbooks already in use within an organisation, DTs may be integrated to provide analysts with new perspectives on the incident. This process is shown in Figure 1.

The first step towards integrating the use of DTs is to examine the activities that are performed at each stage in the playbook. By examining the activities and the types of questions that are asked of the real-world system, it is then possible to identify the most-relevant form of DT assistance. Each activity should be classed according to the areas of support outlined in Table 3. The aim is to determine whether or not the main purpose of the activity is:

(1) to document or communicate to another stakeholder;
(2) to assess the current state of the system; or
(3) to intervene in (or modify) the operation of the system.

If the main objective is to document or communicate to another stakeholder, then there is opportunity for an auto-documentation or communications feature to be implemented to aid the human investigator in documentation of the investigation and communications with other stakeholders, such as plant operators, who manage the physical process.

If the main objective is to assess the current state of the system, then there is opportunity to explore the use of replicated systems, high-fidelity physics-based simulations, and machine learning methods, such as recurrent neural networks, to try and view how the system *should* behave, and use these models to understand the current state of the system.

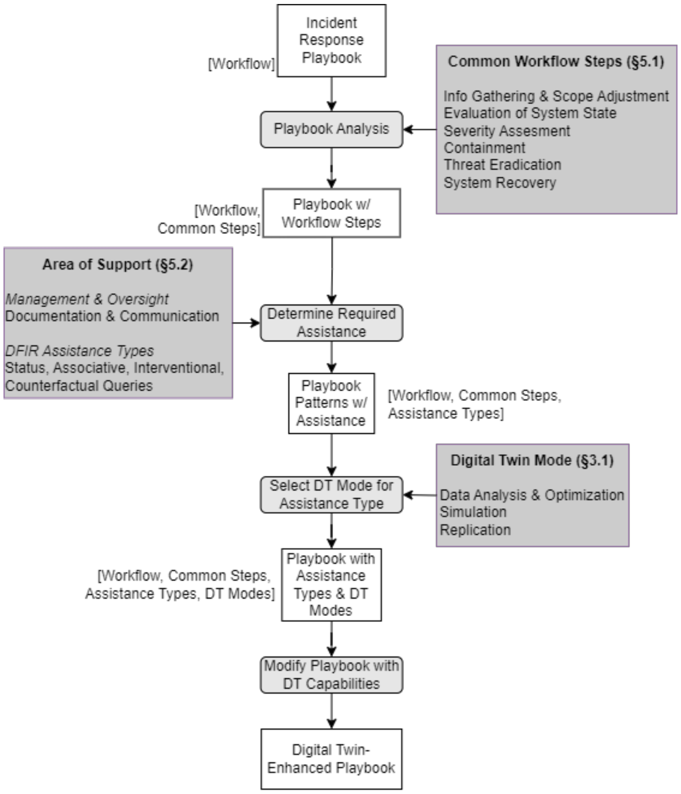
Incident Response Playbook
[Workflow]
Common Workflow Steps (§5.1)
Info Gathering & Scope Adjustment
Evaluation of System State
Severity Assesment
Containment
Threat Eradication
System Recovery
Playbook Analysis
[Workflow, Common Steps]
Playbook w/ Workflow Steps
Area of Support (§5.2)
Management & Oversight
Documentation & Communication
DFIR Assistance Types
Status, Associative, Interventional, Counterfactual Queries
Determine Required Assistance
Playbook Patterns w/ Assistance
[Workflow, Common Steps, Assistance Types]
Digital Twin Mode (§3.1)
Data Analysis & Optimization
Simulation
Replication
Select DT Mode for Assistance Type
[Workflow, Common Steps, Assistance Types, DT Modes]
Playbook with Assistance Types & DT Modes
Modify Playbook with DT Capabilities
Digital Twin-Enhanced Playbook

Figure 1: The process for creating DT-enhanced incident response workflows.

If the main objective is to intervene or interfere in the operation of the system, then there is opportunity to explore the use of high-fidelity simulations, for example, to model potential effects of making changes to the system configurations (intervening) *before* they are implemented in the real-world system.

6 USE CASE: A COMPROMISED PLC PLAYBOOK

This section presents a series of technical vignettes from a playbook investigating a potential Programmable Logic Controller (PLC) reprogramming attack in a critical infrastructure[1] and demonstrates how our approach can be applied to enhance its execution with DTs. An assumption made for this use case is that the attacker a)

[1] The full workflow will be made available upon publication.

has access to the network [23], and b) is able to capture and replay modified packets to the PLC [3]. These vignettes are adapted from a playbook workflow that we designed after performing the analysis detailed in Section 5.1. The assistance offered by DTs at each stage is then described.

The initiating condition for the workflow is a SIEM alert that detects an unauthorised connection to a PLC. For the sake of brevity, processes that share common goals or concepts have been grouped together. Furthermore, the initiating condition has been omitted from the vignettes, but it should be noted that (as documented in Table 1) DTs can be used for identification of anomalous conditions that trigger IR playbooks. The following subsection are organized based on the analysis that is described in Section 5.1.

6.1 Information Gathering, Scope Adjustment

At the beginning of the workflow, information is gathered from different sources and analysed to understand the scope of the attack and subsequently reveal an initial level of severity of the incident (although this does have a dedicated process in the workflow – see Section 6.3). In this context, consider the following step-by-step workflow instructions:

(1) Gather network data (.pcaps, firewall logs, PLC logs)
(2) Determine what subsystem is affected
(3) Enumerate affected systems

The information gathering stage of the workflow is considered a *Status Query*. Data-driven analysis, simulations, or replicated systems can be used to perform status queries.

SIEM solutions combine the evidence in (1) but the selection or filtering of related data remains a largely manual task. While SOAR aims to automate this task (based on existing SIEM alert information, for example), a DT containing a virtual copy of an organisation's infrastructure can automatically supply the host information. Cyber-attacks can be executed in the DT to find the difference between an uncompromised configuration of the infrastructure and one which has been attacked. The documented difference can be used to confirm indicators of compromise in the real-world system.

6.2 Evaluating System State

Throughout the workflow, there are activities that require the investigator to consider the evidence and judge the perceived anomaly as malicious or benign. When investigating a potential unauthorised PLC reprogramming attack, passive approaches may not provide enough information to make this judgement. Therefore, it is necessary to directly connect to the PLC and analyse the configuration. In this context, consider the following step-by-step workflow instructions:

(1) Log into a machine with access to PLC (e.g. Engineering Workstation)
(2) Load known uncompromised PLC configuration
(3) "Go online" (connect to) to PLC
(4) Compare configurations and determine if PLC has been altered

This stage of the workflow is considered a *Status Query* as it aims to obtain the status of the system. As a result, all MOs can be applied here. Passive approaches used to determine the PLC's state, such as analysing network data, could be considered an *Associative Query*, as the goal is to establish an association between network data and PLC state. For example, finding evidence that the PLC responded positively to an unauthorised connection would lead an investigator to believe that the PLC has a higher chance of being compromised. Data-driven analysis methods can provide benefit here, with the arguable advantage of requiring less effort to implement due to advancements in machine learning, for example.

If the reconfiguration of the PLC affects its behaviour (i.e. control), then data-driven DTs that are built to model the uncompromised PLC can be used as a benchmark for determining if the PLC has been altered. This form of DT can run alongside the PLC and provide this service before the step-by-step instructions need to be executed in the workflow.

In (2), replication- and simulation-based DTs can provide two different views on the PLC state. Firstly, a replication- or simulation-based DT of the PLC can be built from system specification, including PLC configuration (control logic, hardware configuration, etc.) to closely mimic or emulate the intended PLC behaviour and state.

A second replication-based DT of the PLC can be implemented to mirror the PLC's inputs, outputs, configurations (including malicious alterations) and therefore exist as a copy of the real PLC's system state.

These two approaches to PLC twinning (1-to-1 mirroring/copying, and *set-and-forget* simulation/replication) can provide two ready-to-view sources of information on the state of the PLC for the SOC operator. Furthermore, with a pre-approved implementation of these PLC twins, it may remove the need for requesting permission from operations staff.

6.3 Incident Severity Assessment

Incident severity is perhaps more important in CPSs than in IT environments, due to the potential impact of a successful cyber-attack. Due to this increased risk, at times it is necessary to escalate issues to operations personnel that monitor the physical process. This escalation is taken immediately if there is risk of a safety-related event (physical damage to equipment, release of radioactive/toxic material into the environment, etc.). DTs have the potential to automatically evaluate the incident severity and even begin forecasting future events. In this context, consider the following step-by-step workflow instructions:

(1) Do affected systems have potential to compromise safety?
   – YES: Escalate to operations personnel immediately.
   – NO: GO TO (2).
(2) Can SOC access and eradicate threat without accessing control network(s)?
   – YES: GO TO (3).
   – NO: Contact operations personnel to arrange access.
(3) No safety concerns, no access to control networks required: Continue playbook as normal.

The establishment of associations between what has been seen and what could occur as a result is an *Associative Query*. Again, DTs in the data analysis mode of operation may provide the most efficient results. However, incident severity assessment may also entail *Interventional* queries to understand the best or worst possible

Table 4: The main forms of query required, and the dominant mode of operation (MO) at each stage of the incident response playbook explored in the use case. *D* = Data Analysis and Optimisation, *S* = Simulation, and *R* = Replication.

	Info. Gathering, Scope Adjust.	Evaluating System State	Incident Severity Assessment	Containment	Threat Eradication	System Recovery
Status	✓	✓	x	x	x	✓
Associative	x	✓	✓	x	x	x
Interventional	x	x	✓	✓	✓	x
Counterfactual	x	x	x	x	x	x
MO	D, S, R	D, S, R	D, S	S, R	R, S	D, S

outcomes, as well as the limits of the operator's ability to control the situation. Simulations yield the best results in this case.

Incident severity assessment can take many forms. The potential impact of a cyber-attack to a given system is based on the type and scope of attack, and should be already defined as part of an IR plan (see *Preparation* in [22]). Incident severity guidelines may be offered by national authorities or regulators [20].

DTs can automatically gather information about the scope of the attack and perform an explainable assessment of the incident severity level. For example, SOC operators investigating security events at the process control level can use a dedicated DT interface to select the affected hosts and type of attack under investigation.

As the investigation continues and investigators adjust the perceived scope of the attack, the DT can apply severity assessment guidelines to the information provided. Furthermore, DT simulations of *worst-case scenario* events can be used to provide more concrete evidence of cascading effects, for example.

6.4 Containment

Some short-term containment strategies can be taken swiftly to effectively contain a cyber-attack to a single host or network. Some containment strategies, such as blocking IP addresses or disconnecting or shutting down hosts, directly affect the availability of the now-isolated host or service. OT environments, which have strict availability requirements for the continued operation of heterogeneous systems including diverse, interdependent, interconnected services and data flows, require more assessment of the potential consequences of these system interventions. In this context, consider the following step-by-step workflow instructions:

(1) Does the malicious traffic originate from a documented machine or known host?
   - YES: GO TO (2).
   - NO: GO TO (3).

(2) Does the malicious traffic originate from a machine that performs a critical function?
   - YES: Escalate to operations personnel for containment and remediation assistance.
   - NO: GO TO (3).

(3) Block IP address. Document changes on security tooling.

At its core, containment is intervention to prevent an issue from spreading. *Interventional* reasoning, e.g. *what if we turn off this system*, is the main area of support offered by DTs at this stage in the workflow. Simulations and replicated systems can be used to determine the best strategies for approaching the containment of the problem.

Questions (1) and (2) can be answered by virtue of a DT platform being built upon specification, documentation, and expert knowledge of the system. The DT platform will contain information on known hosts and their system criticality. Answers to (1) and (2) can therefore not only be automated or suggested to the end user but also explained by producing the relevant host information.

Action taken in (3) may result in unknown adverse affects, therefore edits to firewall rules, IDS/IPS configurations, etc. may be executed in a DT before being implemented on the real-world system.

6.5 Threat Eradication

Once the source host of the attack has been found and contained, the eradication of the root cause requires further analysis, and ultimately a direct system intervention to terminate the malicious processes or software. In this context, consider the following step-by-step workflow instructions:

(1) Examine processes and applications on affected host
(2) Compare to baseline services in a machine's documentation
(3) Terminate extra or unknown processes
(4) Investigate suspicious processes that are part of system baseline

Similar to the containment phase, *Interventional* reasoning is the main area of support offered by DTs. As a result, simulation and – in this example – replication-based DTs offer the best assistance.

Examining and terminating processes on a host in an IT environment is a common occurrence in most industries; however, hosts in OT environments that interact with other OT components can perform critical functions that require high system availability. A passive approach to IR is therefore favoured; i.e. to investigate a cyber-attack with as little interaction with the real-components as possible.

With eradication, however, it is inevitably required to address the source of the malicious activity, which in this vignette means examining a host in an OT environment. DTs can serve a baseline for (1) and (2) to determine how the host system *should* look and behave. Investigations on hosts can be aided by virtualisation technologies, allowing part of the investigation to take place virtually, reducing the time spent on a possibly fragile host or OT network.

6.6 System Recovery

Once a threat has been eradicated and stakeholders are confident that the compromised host can be brought back online and introduced back into the system, e.g. unblocking IP, physically connecting the host again, etc., there is a period of time dedicated to closely monitoring the previously compromised systems. Data-driven DTs that implement machine learning methods can be used to mimic the physical process and run in parallel with the real-world system to determine if the system is behaving as it should. Simulation-based DTs can execute similarly with the added benefit of being able to change parameters to answer "what-if" questions.

6.7 Discussion

This section has illustrated how our approach to identifying potential applications and modes of operation of DTs to enhance IR playbooks can be used. The example playbook workflow has characteristics that are typical of those described in Section 5.1. Table 4 summarizes the different types of queries that are pertinent during playbook execution and the proposed DT modes of operation to support those queries. It can be seen that counterfactual queries are not used during the execution of the workflow steps that we have described – these are used during the *Preparation* and *Lessons Learned* phases. Counterfactual queries can be best supported by simulations-based DTs and replication-based DTs. During post-incident analysis, these DTs enable organisations to analyse how the incident *could have* played out. This can also be used to train staff to prepare for similar scenarios in the future.

This example can be used to illustrate a practical application of our approach. It may be that via security assurance activities, such as a cyber security exercise, it has been identified there are bottlenecks or issues with the execution of parts of a playbook – areas where there is room for improvement. Our approach can be used to identify where DTs of which specific systems could be used to make improvements and the pertinent modes of operation that are needed for specific types of assistance. For example, it may be that it has been determined that concerns surrounding the potential adverse effects (risks) of containment of a PLC resulted in unacceptable service restoration times. Using our approach, it can be determined that a playbook could be enhanced with a DT of the target system that is capable of either simulating or replicating its behaviour in order to perform interventional queries that relate to containment activities, e.g., *what happens to a controlled process if I block the IP address of a compromised PLC?* The intention is that the DT can assist in determining the answers to these kinds of queries in order for stakeholders (security analysists, operators, etc.) to make more rapid and correct decisions.

7 CONCLUSION

Our paper illustrates how DTs can assist IR playbook execution, provide stakeholder unique, more-complete views on system state, and ultimately help users make better informed decisions. To achieve this, we propose an approach for analysing existing IR workflows and applying the most appropriate form of DT. Our analysis reveals that data-driven DTs have the potential to assist in the identification of anomalies and system state monitoring, while activities with system intervention, such as containment and eradication, are best assisted with simulations and replicated systems.

From our analysis of existing incident response procedures, we find that IT-focused IR playbooks rarely ask interventional ("what-if") questions, due to the focus on confidentiality and integrity during IR. In contrast, the physicals operations domain concerns itself with availability, due to the critical nature of the systems. As a result, the development of abnormal and emergency operating procedures (AOPs, EOPs) includes the evaluation of the "what-if" scenarios *in advance*, resulting in conservative (i.e. safety-first), inflexible procedures.

IR in OT environments needs to provide a balance between timely remediation and appropriate levels of system interference. Therefore, there must be interventional reasoning when evaluating IR activities. Unlike the IT domain, it is not possible to simply unplug machines, and it might not be possible to get (immediate) access to systems or data to perform an investigation. Our paper demonstrates how DTs can provide the tools to support the evaluation of these interventional queries, and how they can be integrated into existing playbooks.

For future work, we will explore *automated* DT response strategies, and how they can influence playbooks at the design stage through parallel execution and information prioritisation.

REFERENCES

[1] David Allison, Paul Smith, and Kieran McLaughlin. 2022. Digital Twin-Enhanced Methodology for Training Edge-Based Models for Cyber Security Applications. In *2022 IEEE 20th International Conference on Industrial Informatics (INDIN)*. IEEE, Perth, Australia.

[2] Manolya Atalay and Pelin Angin. 2020. A Digital Twins Approach to Smart Grid Security Testing and Standardization. In *2020 IEEE International Workshop on Metrology for Industry 4.0 & IoT*. 435–440. https://doi.org/10.1109/MetroInd4.0IoT48571.2020.9138264

[3] E. Biham, S. Bitan, Aviad Carmel, Alon Dankner, Uriel Malin, and A. Wool. 2019. Rogue 7 : Rogue Engineering-Station Attacks on S 7 Simatic PLCs.

[4] Adrien Bécue, Eva Maia, Linda Feeken, Philipp Borchers, and Isabel Praça. 2020. A New Concept of Digital Twin Supporting Optimization and Resilience of Factories of the Future. *Applied Sciences* 10, 13 (Jan. 2020), 4482. https://doi.org/10.3390/app10134482 Number: 13 Publisher: Multidisciplinary Digital Publishing Institute.

[5] Paul Cichonski, Tom Millar, Tim Grance, and Karen Scarfone. 2012. *Computer Security Incident Handling Guide : Recommendations of the National Institute of Standards and Technology*. Technical Report. National Institute of Standards and Technology. NIST SP 800–61r2 pages. https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-61r2.pdf DOI: 10.6028/NIST.SP.800-61r2.

[6] Incident Response Consortium. 2023. Incident Response Playbooks Gallery. https://www.incidentresponse.org/playbooks/

[7] Cybersecurity and Infrastructure Security Agency (CISA). 2022. *Operational Procedures for Planning and Conducting Cybersecurity Incident and Vulnerability Response Activities in FCEB Information Systems*. Technical Report. Cybersecurity and Infrastructure Security Agency (CISA), Arlington, VA. https://www.cisa.gov/uscert/sites/default/files/publications/federal-government-cybersecurity-incident-and-vulnerability-response-playbooks-508c.pdf

[8] Marietheres Dietz, Ludwig Englbrecht, and Günther Pernul. 2021. Enhancing Industrial Control System Forensics using Replication Based Digital Twins. In *Advances in Digital Forensics XVII*, Gilbert Peterson and Sujeet Shenoi (Eds.). Springer International Publishing, Cham, 21–38.

[9] Marietheres Dietz and Gunther Pernul. 2020. Unleashing the Digital Twin's Potential for ICS Security. *IEEE Security Privacy* 18, 4 (July 2020), 20–27. https://doi.org/10.1109/MSEC.2019.2961650

[10] Marietheres Dietz, Benedikt Putz, and Günther Pernul. 2019. A Distributed Ledger Approach to Digital Twin Secure Data Sharing. In *Data and Applications Security and Privacy XXXIII (Lecture Notes in Computer Science)*, Simon N. Foley (Ed.). Springer International Publishing, Cham, 281–300. https://doi.org/10.1007/978-3-030-22479-0_15

[11] Marietheres Dietz, Manfred Vielberth, and Günther Pernul. 2020. 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, 1–9. https://doi.org/10.1145/3407023.3407039

[12] Matthias Eckhart and Andreas Ekelhart. 2018. A Specification-Based State Replication Approach for Digital Twins. In *Proceedings of the 2018 Workshop on Cyber-Physical Systems Security and PrivaCy - CPS-SPC '18*. ACM Press, Toronto, Canada, 36–47. https://doi.org/10.1145/3264888.3264892

[13] Matthias Eckhart and Andreas Ekelhart. 2018. Towards Security-Aware Virtual Environments for Digital Twins. In *Proceedings of the 4th ACM Workshop on Cyber-Physical System Security - CPSS '18*. ACM Press, Incheon, Republic of Korea, 61–72. https://doi.org/10.1145/3198458.3198464

[14] Matthias Eckhart and Andreas Ekelhart. 2019. Digital Twins for Cyber-Physical Systems Security: State of the Art and Outlook. In *Security and Quality in Cyber-Physical Systems Engineering*. Springer, 383–412. https://doi.org/10.1007/978-3-030-25312-7_14

[15] Matthias Eckhart, Andreas Ekelhart, and Edgar Weippl. 2019. Enhancing Cyber Situational Awareness for Cyber-Physical Systems through Digital Twins. In *2019 24th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*. 1222–1225. https://doi.org/10.1109/ETFA.2019.8869197

[16] Philip Empl, Daniel Schlette, Daniel Zupfer, and Günther Pernul. 2022. SOAR4IoT: Securing IoT Assets with Digital Twins. In *Proceedings of the 17th International Conference on Availability, Reliability and Security*. ACM, Vienna Austria. https://doi.org/10.1145/3538969.3538975

[17] Integrated Adaptive Cyber Defense (IACD). 2017. A Specification for Defining, Building and Employing Playbooks to Enable Cybersecurity Integration and Automation. https://tinyurl.com/IACDSpec

[18] International Atomic Energy Agency. 1985. *IAEA-TECDOC-341 - Developments in the Preparation of Operating Procedures for Emergency Conditions of Nuclear Power Plants*. Technical Report. International Atomic Energy Agency, Vienna, Austria.

[19] International Atomic Energy Agency. 2006. *Development and Review of Plant Specific Emergency Operating Procedures*. Technical Report 48. International Atomic Energy Agency, Vienna, Austria. 103 pages.

[20] International Atomic Energy Agency. 2016. *Computer Security Incident Response Planning at Nuclear Facilities*. Technical Report IAEA-TDL-005. International Atomic Energy Agency, Vienna, Austria. https://www-pub.iaea.org/MTCD/publications/PDF/TDL005web.pdf

[21] Blake Johnson, Dan Caban, Marina Krotofil, Dan Scali, Nathan Brubaker, and Christopher Glyer. 2017. Attackers Deploy New ICS Attack Framework "TRITON" and Cause Operational Disruption to Critical Infrastructure. https://www.fireeye.com/blog/threat-research/2017/12/attackers-deploy-new-ics-attack-framework-triton.html [Online; accessed 2018-05-25].

[22] Patrick Kral. 2021. *The Incident Handler's Handbook*. Technical Report. Escal Institute of Advanced Technologies (SANS Institute), Rockville, Maryland, United States.

[23] Robert Lee, Michael Assante, and Tim Conway. 2016. *Analysis of the Cyber Attack on the Ukrainian Power Grid*. Technical Report. Electricity Information Sharing and Analysis Center (E-ISAC), Washington D.C., USA. https://nsarchive.gwu.edu/sites/default/files/documents/3891751/SANS-and-Electricity-Information-Sharing-and.pdf

[24] Microsoft Corporation. 2022. Microsoft Incident Response Playbooks. https://learn.microsoft.com/en-us/security/compass/incident-response-playbooks

[25] Judea Pearl. 2019. The Seven Tools of Causal Inference, with Reflections on Machine Learning. *Commun. ACM* 62, 3 (Feb. 2019), 54–60. https://doi.org/10.1145/3241036

[26] Paula Savioja, Leena Norros, Leena Salo, and Iina Aaltonen. 2014. Identifying resilience in proceduralised accident management activity of NPP operating crews. *Safety Science* 68 (Oct. 2014), 258–274. https://doi.org/10.1016/j.ssci.2014.04.008

[27] Manfred Vielberth, Magdalena Glas, Marietheres Dietz, Stylianos Karagiannis, Emmanouil Magkos, and Günther Pernul. 2021. A Digital Twin-Based Cyber Range for SOC Analysts. In *Data and Applications Security and Privacy XXXV (Lecture Notes in Computer Science)*, Ken Barker and Kambiz Ghazinour (Eds.). Springer International Publishing, Cham, 293–311. https://doi.org/10.1007/978-3-030-81242-3_17

[28] Kim Zetter. 2014. *Countdown to Zero Day: Stuxnet and the Launch of the World's First Digital Weapon*. Crown Publishing Group, USA.

Received 01 March 2023