POLITECNICO DI MILANO
DIPARTIMENTO DI ENERGIA

Doctoral Programme in Energy and Nuclear Science and Technology

XXX Cycle (2014-2018)

A Modeling and Simulation Framework for Analyzing Failures in Cyber-Physical Systems for Energy Applications

(The Ph.D. Thesis)

Doctoral Dissertation of:	Wei Wang	
Supervisors:	Dr.	Francesco Di Maio
	Prof.	Enrico Zio
Tutor:	Dr.	Francesco Di Maio
Coordinator:	Prof.	Carlo Enrico Bottani

13 December 2018, Milano

This thesis is dedicated to my mother and my wife
who are the two most important women in my life,
for their love, endless support, and encouragement.

Questa tesi è dedicata a mia madre e mia moglie
che sono le due donne più importanti della mia vita,
per il loro amore, supporto infinito e incoraggiamento.

谨以此论文献给我的母亲和我的妻子
——我人生中最重要的两位女性，
感谢她们对我永恒的爱、支持与鼓励。

[This page intentionally left blank.]

ACKNOWLEDGEMENTS

RINGRAZIAMENTI
致谢

**** to my supervisors and my Ph.D. programme ****
**** ai miei supervisori e al mio programma di dottorato ****

I would like to express my special appreciation and thanks to my supervisors: Dr. Francesco Di Maio and Prof. Enrico Zio, for their exemplary guidance, great technical supports, numberless insightful discussions, timeless patience, and constant encouragements throughout all the course of this thesis. The blessing, help and guidance given by them time to time shall carry me a long way in the journey of my future research life. It is a pleasure and a great honor to pursue a Ph.D. degree under their supervision.

I would like to express my sincere appreciation and gratitude to the Ph.D. programme "Energy and Nuclear Science and Technology" in Department of Energy, Politecnico di Milano, and to the program coordinator, Prof. Carlo Enrico Bottani, for a four-year hosting.

I would like to thank China Scholarship Council and the Education Office of Embassy of P. R. China in Italy for awarding me the scholarship under the support of State Scholarship Fund, allowing me to pursue the 48-month study in Politecnico di Milano as Ph.D. student.

Vorrei esprimere uno speciale ringraziamento ai miei supervisori, il Dr. Francesco Di Maio e il Prof. Enrico Zio, per la loro guida esemplare, l'enorme supporto tecnico, i confronti numerosi e profondi, la pazienza instancabile, e gli incoraggiamenti constanti durante l'intero corso di questa tesi. Il sostegno, l'aiuto e la guida che mi hanno continuamente dato mi accompagneranno ancora a lungo nel viaggio della mia futura vita nella ricerca. È un piacere e un onore perseguire il titolo di dottorato con la loro guida.

Vorrei esprimere la mia sincera gratitudine al programma di dottorato "Science e tecnologie energetiche e nucleari" presso il Dipartimento di Energia del Politecnico di Milano, e al coordinatore del programma, il Prof. Carlo Bottani, per i quattro anni di ospitalità.

Vorrei ringraziare il China Scholarship Council e l'Ufficio Istruzione dell'Ambasciata della Repubblica Popolare Cinese in Italia per avermi accordato, col supporto dello State Scholarship Fund, la borsa che mi ha permesso di spendere 48 mesi di studio come studente di dottrato presso il Politecnico di Milano.

I am grateful to the two anonymous reviewers and the defense committee members, whose comments have allowed improving the thesis significantly.

**** to my family ****
**** 之于我的家人 ****

I would like to offer my sincerest thanks to my parents for their unconditional and everlasting love, support and encouragement.

I would like to offer my most reverent thanks to my parents-in-law for their trust, broad-minded comfort and support.

I would like to offer my most affectionate thanks to my wife for her deep love, endless support and encouragement.

I would like to offer my special thanks to my sister, my brother-in-law and my twin nephews for their quiet contributions to the family when I left home for persuing my study.

I would also like to offer my heartfelt thanks to the other family members for having supported and appreciated my pursuing of the doctoral degree.

我要向我的父母表达最真挚的感谢，感谢他们对我养育之恩，感谢他们对我无条件的、永恒的爱、关心和照顾，才使我安心学习和研究，顺利完成学业！

我要向我的岳父母表达最尊敬的感谢，感谢他们一直以来对我的信任、宽慰和包容，感谢他们给予我和妻子爱情的支持，才使我集中精力，投入科研！

我要向我的妻子表达最深情的感谢，感谢她对我深情、温暖的爱与陪伴。拾年如初，汝复芬葩焕耀；愿，自今已往，执子期岁月静好！

我要向我的姐姐、姐夫及双胞胎外甥表达最特别的感谢，感谢他们在我离家求学的这些年对家庭的默默付出与贡献！

我要向我的其他家人表达衷心的感谢，在我追求博士学位的过程中，感谢他们一直以来的照顾与欣赏！

**** to my friends and colleagues ****

I would also like to take this opportunity to thank all my friends and colleagues in the LASAR (Laboratory of signal and risk analysis) research group: Prof. Piero Baraldi, Dr. Michele Compare, Yang Hu, Sameer Al-Dahidi, Marco Rigamonti, Francesco Cannarile, Davide Colli, Edoardo Tosoni, Alessandro Mancuso, Shiyu Chen, Zhe Yang, Rodrigo Mena, Seyed Mojtaba Hoseyni, Mingjing Xu, Federico Antonello, Luca Bellani, Ahmed Shokry, Dario Valcamonico, Xuefei Lu, and Zhaojun Hao. It has been a great pleasure to work with them during the last four years. I want to give them my great thanks for everything they have done, for their encouragement and their disposition to offer me a hand whenever I needed help.

I am deeply grateful to my friends and colleagues, Profs. Jiejuan Tong and Jun Zhao of Tsinghua University, Prof. Antonio Cammi and Dr. Stefano Lorenzi of Politecnico di Milano, Sai Zhang of University of Illinois at Urbana Champaign, Yue Zhao of China Securities Co., Ltd., Yue Pi and Yun Liu of China Nuclear Power Engineering Co., Ltd., Hao Wu of Wuhan Secondary Ship Design and Research Institute, Xiaolei Pan of Shanghai University of Electric Power, Dacheng Zhang of Kunming University of Science and Technology, Yunfei Zhao of The Ohio State University, Zhiguo Zeng of Université Paris Saclay, Youhu Zhao and Taotao Zhou of Huawei Technologies Co., Ltd., Yu Yu of North China Electric Power University, Yanhua Zou of Hunan Institute of Technology, Xiaoyu Cui and Tongyang Li of Beihang University, Yu Miao of Beijing Jiaotong University, for their discussions and collaborations. In addition, I would like to thank all my friends in life, for their warm helps and supports from all aspects.

Specially, I am deeply grateful to Edoardo Tosoni, Dario Valcamonico and Dr. Francesco Di Maio, for their Italian translation efforts that have greatly contributed to the composition of this thesis.

Specialmente, sono profondamente grato a Edoardo Tosoni, Dario Valcamonico e Dr. Francesco Di Maio, i cui sforzi nelle traduzioni in italiano hanno contribuito grandemente alla preparazione di questa tesi.

ABSTRACT

Cyber-Physical Systems (CPSs) feature a tight combination of (and coordination between) physical processes and cyber systems, for automation and control. The integration of cyber resources into energy production processes enables the energy CPSs to be real-time monitored and dynamically controlled, during normal operation as well as in case of accidents. Specifically to nuclear energy, the introduction of digital Instrumentation and Control (I&C) systems allows Nuclear Power Plants (NPPs) to take advantage of CPSs, for improved process monitoring, control, and protection.

CPSs are subjected to failures (i.e., deviations from the system expected behaviors, which can lead the system to damage) due to degradations and failures of the physical components, and to intentional or accidental breaches in the cyber security. Thus, CPSs failure analysis must comprise safety and security aspects.

The objective of the Ph.D. work is to develop a general modeling and simulation framework for the failure analysis of CPSs, which include I. identification and prioritization of hazards and threats (to identify the conditions that trigger anomalies in the systems and their causes), II. failure scenarios modeling and simulation (to characterize the system behavior under different operational conditions, including hazardous and malicious ones), III. consequence analysis (to explore the effects of stochastic component failures and cyber attacks onto the CPS functionality) and, IV. protection design (to take decisions on recovery measures for increasing system resilience). The proposed framework is fundamental to address all possible hazards and threats in a comprehensive and holistic way.

With respect to I, II and III, the framework addresses the tasks keeping hazards and threats (with their effects on the CPSs functionalities) separate: on one hand, the hazards (i.e., the stochastic failures more affecting the CPS safety) are proposed to be identified by the Multi-State

Physical Modeling (MSPM) approach that accounts for uncertainties on environmental conditions, aging and degradation of component failure events, whereas the threats (i.e., the malicious attacks more affecting the CPS security) by a Monte Carlo (MC)-based exploration framework that, based on safety margin estimation, allows simulating the effects of the cyber threats on the system functionality and prioritizing the most vulnerable components of the CPSs.

With respect to the IV, to protect the CPS from (unknown and uncertain) cyber attacks, we propose an Adversarial Risk Analysis (ARA) approach to provide a novel one-sided prescriptive support strategy for the defender to optimize the defensive resource allocation, based on a subjective expected utility model. Once hazards and threats are well identified and protections selected, efforts can be devoted to optimally design protections. Moreover, the prompt recognition and distinction of cyber attacks from component failures in CPSs rely on the simultaneous treatment, within a consolidated Non-Parametric Cumulative Sum (NP-CUSUM) approach, of the measurements taken from redundant channels.

Specifically, case studies considered include nuclear CPSs (i.e., a typical Reactor Protection System (RPS) of NPPs, and the digital I&C system of an Advanced Lead-cooled Fast Reactor European Demonstrator (ALFRED)).

KEYWORDS

Cyber-Physical System; Failure Analysis; Hazards; Threats; Component Failure; Cyber Attack; Uncertainty; Monte Carlo; Reliability Assessment; Cyber Security; Multi-State Physics Modeling; Safety Margin; Non-Parametric Cumulative Sum; Adversarial Risk Analysis; Digital Instrumentation and Control (I&C) System; Reactor Protection System; Nuclear Power Plant; Advanced Lead-cooled Fast Reactor European Demonstrator.

SOMMARIO

I sistemi cyber-fisici (CPSs) sono caratterizzati da una stretta combinazione di (e coordinazione tra) processi fisici e sistemi cyber. L'integrazione di risorse cyber nei processi di produzione di energia consentono il controllo dinamico e il monitaraggio in tempo reale dei CPSs, sia nelle normali operazioni che in caso di incidenti. Specificamente per l'energia nucleare, l'introduzione di sistemi di Strumentazione e Controllo (I&C) digitali rende le Centrali Nucleari (NPPs) equivalenti ad un sistems cyber-fisico, garantendo un migliore monitoraggio del processo, controllo e protezione.

I CPSs sono soggetti a guasti (i.e. deviazioni dal comportamento atteso del sistema, che possono danneggiarlo) a causa del degrado e rottura dei componenti fisici del sistema, e di attacchi intenzionali o accidentali ai componenti cyber. Dunque, l'analisi dei guasti dei sistemi cyber-fisici deve considerare la sicurezza di entrambi gli aspetti.

L'obiettivo del lavoro di Ph.D. è lo sviluppo di un framework generale di modellizzazione e simulazione per l'analisi dei guasti dei CPSs, che include I. l'identificazione e la prioritizzazione delle minacce e delle occorrenze incidentali (per identificare le condizioni che comportano anomalie nei sistemi e le loro cause), II. La modellizzazione e la simulazione degli scenari incidentali di guasto (per caratterizzare il comportamento del sistema in diverse condizioni operative, incluse condizioni di incidente e pericolo), III. L'analisi delle conseguenze (per esplorare gli effetti di fallimento stocastico dei componenti e attacchi cyber sulla funzionalità del CPS) e, IV. design per la protezione (per prendere decisioni su misure di recupero per migliorare la resilienza del sistema). Il framework proposto è fondamentale per considerare in maniera completa e olistica tutti gli eventi incidentali e le minacce al sistema.

Con riferimento a I, II e III, nel framework proposto eventi incidentali e minacce (e i loro effetti sulla funzionalità dei CPSs) vengono mantenuti separati: da un lato, gli eventi incidentali (i.e. i fallimenti stocastici incidentali del sistema e dei suoi componenti) sono identificati con un approccio Multi-State Physical Modeling (MSPM) che include incertezze sulle condizioni ambientali e gli eventi di fallimento dovuti al degrado e invecchiamento dei componenti; dall'altro le minacce (i.e. gli attacchi malevoli e intenzionali alla sicurezza del sistema) sono identificate da uno schema di esplorazione Monte Carlo (MC) che, basato sulla stima dei margini di sicurezza probabilistici, consente la simulazione delle conseguenze delle minacce cyber sulla funzionalità del sistema e prioritizza i componenti più vulnerabili dei CPSs.

Con riferimento a IV, per proteggere il sistema da attacchi cyber (sconosciuti e incerti), viene proposto un approccio Adversarial Risk Analysis (ARA) per fornire al difensore del CPS un'innovativa strategia per ottimizzare l'allocazione di risorse di difesa. Identificati correttamente gli eventi incidentali e le minacce, e selezionate le misure di protezione, è possibile concentrare gli sforzi per ottimizzare il design delle misure di riconscimento delle minacce cyber a cui il CPS è soggetto: il tempestivo riconoscimento e distinzione di attacchi cyber dal fallimento dei componenti nei CPSs proposto nell'ultima parte della tesi è basato sul processamento di misurazioni acquisite da canali ridondanti di monitoraggio del sistema, con approccio Non-Parametric Cumulative Sum (NP-CUSUM).

Specificamente, i casi studio considerati includono sistemi CPSs nucleari (i.e., un tipico Sistema di Protezione del Reattore (RPS) di un impianto nucleare, e il sistema digitale I&C di Advanced Lead-cooled Fast Reactor European Demonstrator (ALFRED)).

PAROLE CHIAVE

Sistema Cyber-Fisico; Cyber Security; Multi-State Physics Modeling; Safety Margin; Non-Parametric Cumulative Sum; Adversarial Risk Analysis; Digital Instrumentation and Control (I&C) System; Sistema di Protezione del Reattore; Centrale Nucleare; Advanced Lead-cooled Fast Reactor European Demonstrator.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

[This page intentionally left blank.]

ABBREVIATIONS

ALFRED	Advanced Lead-cooled Fast Reactor European Demonstrator
ARA	Adversarial Risk Analysis
BBN	Bayesian Belief Network
BPL	Bistable Processor Logic
CDF	Cumulative Distribution Function
CPD	Conditional Probability Distribution
CPT	Conditional Probablity Table
CPS	Cyber-Physical System
CR	Control Rod
DAC	Digital-to-Analog Converter
DoS	Denial of Service
ESS	Emergency Shutdown Signal
FA	Fuel Assemblies
FAR	False Alarm Rate
FDI	False Data Injection
HMI	Human-Machine Interface
I&C	Instrumentation and Control
IDS	Intrusion Detection System
IP	Internet Protocol
LCL	Local Coincidence Logic
LSB	Least Significant Bit
MC	Monte Carlo
MCM	Markov Chain Model
MSPM	Multi-State Physics Model(ing)
NP-CUSUM	Non-Parametric CUmulative SUM
NPP	Nuclear Power Plant
OS	Order Statistics
PDF	Probability Density Function
PI	Proportional-Integral
PID	Proportional-Integral-Derivative
PSA	Probabilistic Safety Analysis
PSF	Performance Shaping Factor
PTS	Partial Tripping Signal
R&D	Research and Development
RPS	Reactor Protection System
RTB	Reactor Trip Breaker
RTD	Resistance Temperature Detector
SA	Sensitivity Analysis
SCADA	Supervisory Control And Data Acquisition
SG	Steam Generator
SISO	Single Input Single Output

[This page intentionally left blank.]

SECTION I. GENERALITIES

This section of the dissertation describes the context of the Ph.D. research, its relevance, the state-of-the-art methods, the challenges that are addressed, the overview of the developed framework and the description of the industrial applications carried out for the demonstration.

1. INTRODUCTION

Cyber-Physical Systems (CPSs) are supporting the development of our industry and society [1-3]. CPSs feature a tight combination of (and coordination between) the system computational units and physical elements [4-6]. In CPSs, cyber and physical processes are dependent and interact with each other through feedback control loops (e.g., embedded and networked cyber controllers that rely on databases monitor and control with the supervision of operators by user interfaces the system physical variables, whilst physical processes affect, at the same time, the monitoring system and the computation units by wired or wireless networks [4, 5, 7-12]). Besides the benefit of safer operation, the integration of computational resources into physical processes is expected to add new capabilities to stand-alone physical systems by enabling innovative opportunities of connectivity, real-time monitoring, communication, dynamic control and decision support, shifting our business, ways of production processes, controls and services towards new modalities [1, 13].

The transdisciplinary concept of CPS originates in the years 2000s, and is nowadays quite common in aerospace, automotive, transportation, medical and health-care, energy, and other applications [4, 5, 14-16]. Specifically to nuclear energy, the introduction of digital Instrumentation and Control (I&C) systems allows Nuclear Power Plants (NPPs) to take advantage of CPSs [17], for improved process monitoring, control, and protection.

As sketched in Figure 1, the CPS generalizes the traditional term of embedded system that semanticizes the integration of cyber computing resources and the physical world via sensors and actuators in the feedback control loops [4, 5].

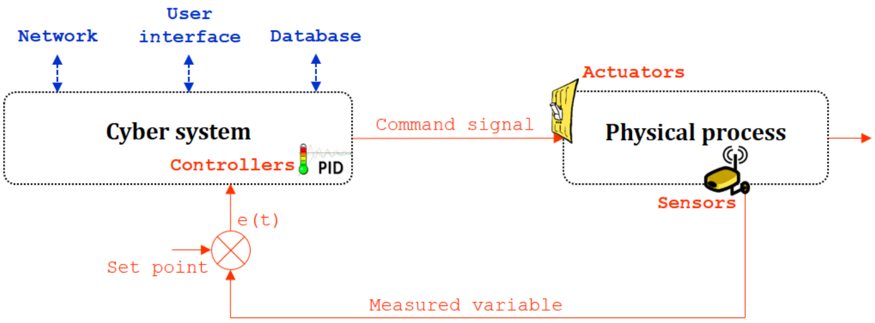


Figure 1 The CPS control scheme

Both physical hazards and cyber threats can compromise the functionality of CPSs and lead to catastrophic consequences, e.g., loss of life and/or revenues [18-21]. For example, on 14 August 2003, the trigger of a software bug (unknown to operators) in the control room alarm system of FirstEnergy Corporation propagated a blackout that brought collapse the entire electric grid along the Northeast United States and portions of Canada for 31 hours, contributing to at least 8 deaths and causing a loss of approximately $6.4 billion [22]; on 23 July 2011, the failure of the automatic train protection module of the train control system due to a lightning storm caused the collision of two trains with 40 people killed and a loss of over $30 million [23]; in 2010, the Natanz nuclear facility in Iran was attacked and infiltrated by a cyber worm, called the Stuxnet, which put the centrifuge out of control for at least 6 months [24]; on August 2006, Unit 3 of the Browns Ferry NPP went into an unwanted shutdown after a flood of malicious data traffic intruded into the plant control system network and blocked the functions of two water recirculation pumps [25].

Risk assessment of CPSs must address both safety and security issues, because not only failures of hardware and software can cause damages and harms, but also cyber attacks can breach the CPS security and lead to serious consequences. Safety concerns stochastic components failures that can result in accidental scenarios leading the system towards unacceptable consequences. Security concerns malicious and intentional attacks that can impair both the physical and cyber parts of the system, and lead to unacceptable consequences.

Traditional risk assessment methodologies have been addressing accidental component failures and software errors, often overlooking the contribution of malicious attacks to the CPSs functionality [26, 27]. However, since CPSs functionality has been shown to be also strongly compromised by security breaches and external attacks, integration of “first principles” attack dynamic models with higher level (CPSs) dynamic features into the existing risk assessment methodologies is needed. CPSs dynamics can be affected by the interrupted communication between the cyber system and the external environment due to human errors injected through user interface, malicious attacks through network systems, and lack of robust databases that are at the basis of digital controllers [28]. Such security attributes are less predictable and depend on many factors (e.g., attacker profile, skills, motivation, etc.), which makes it more difficult for a security analyst to assess and quantify possible scenarios [18]. Thus, if CPS threats and vulnerabilities breaching the cyber security are to be addressed, a confident risk assessment model should be built to address the convergence of safety and security concerns [18, 20, 21].

In this Ph.D. thesis, the objective is to develop a general modeling and simulation framework for the failure analysis of CPSs, which include I. identification of hazards and threats (to identify the conditions that trigger anomalies in the systems and their causes), II. failure scenarios modeling and simulation (to characterize the system behavior under different operational conditions, including hazardous and malicious ones), III. consequence analysis (to explore the effects of stochastic component failures and cyber attacks onto the CPS functionality) and, IV. protection design (to take decisions on recovery measures for increasing system resilience).

The proposed framework can provide results that help to the analysts to identify hazards and threats of CPSs, analyze their causes, model their potential scenarios and consequences, and propose decisions for system protection and resilience. It allows running and analyzing a number of failure scenario simulations including both components stochastic failures and malicious cyber attacks, within which possible (aleatory and epistemic) uncertainties are treated.

The main original contribution of the research lies in the development of a general modeling and simulation framework treating both safety and security aspects of CPSs, that is fundamental to address all possible stochastic failures and cyber threats in a comprehensive and holistic way. Specifically, we have developed novel methods for addressing the following three issues: 1) giving due account to uncertainties affecting aging, degradation and stochastic failures of CPS components, 2) giving due account to the uncertainties that affect threats and vulnerabilities of

CPS to unexpected malicious external attacks, and 3) protection design of CPSs giving due account to cyber attacks and component stochastic failures.

1.1. Cyber Physical Systems

CPSs deeply intertwine the physical and cyber worlds, equip the physical components with cyber attributes (for monitoring and managing of their status) and, thus, network the system at both temporal and spatial scales (for configuring the dynamics processing) [15, 29, 30]. New capability and automation of CPSs change the way we communicate and socially behave and improve the modalities of our living [3]. To realize more improved functionality and service for industrial applications, the CPSs feature the following characteristics:

(1) Real-time computation

CPSs perform intensive real-time tasks including collecting streams of sensor measuring data, computing equations of motion for system dynamics processing and generating actuator commands, to guarantee the timely and efficient system responses within specified time constraints (i.e., deadlines) [7, 31, 32]. In the real-time (digital) signal processing, the CPSs must analyze inputs and generate outputs continuously and simultaneously, but independently of the processing delay [4, 33].

(2) Concurrency and scalability

CPSs require a concurrent treatment of computation for real-time streams of sensors stimuli and actuators controls, improving the effectiveness of the execution in a multiple (large-scale) behavioral system [34-36]. While adding new computational resources into the system functionality, the CPSs have to be scalable of handling the growing amount of systems databases and loads [37].

(3) Stability

To satisfactorily control dynamic systems, it is basically required that changes in the observed behavior of the physical system, in case of small perturbations imposed by the controller, are kept at minimum by CPSs, adapting of control rules on real-time monitoring of physical and environmental variables [4, 38, 39].

(4) Predictability

CPSs have to anticipate the system behaviors, accommodating the stochastic system variability, guaranteeing safety and security during the system operation [5, 40, 41]. In other words, predictability consists in dynamically foreseeing the system behavioral properties, duly treating uncertainty that affects the system dynamic processing [40, 42, 43].

(5) Safety

CPSs prevent intrinsic property degradations to avoid accidental component failures and software errors leading the CPSs to damage and harm [44, 45].

(6) Security

CPSs are resilient to malicious attacks, e.g., cyber attacks, for guaranteeing integrity and confidentiality of information [45, 46]. Cyber-physical security is improved in CPSs by detection of attacks, prompt response, reconfiguration, and restoration of the system functionality [16, 47].

CPS plays an increasingly important role in critical infrastructures and daily life, but also increases safety and cyber security risks. Safety and security problems in the area of CPS have been becoming a global and general issue [3, 48]. Despite diverse CPS architectures, safety and cyber security issues originate from the general principle of CPS of embedding and networking cyber controllers (computation and communication) with physical components (sensors and actuators) in a feedback loop. Consequentially, unintentional components faults or malicious attacks can have severe impact on the CPS functionality and the environment. To mutually and holistically deal with the issue emerging in the whole field of critical infrastructures, it is necessary to develop a modeling and simulation framework for the analysis of failures in general CPSs.

CPSs failure analysis must comprise both safety and security aspects. The dual safety and security share many commonalities. One difficulty lies in the fact that hazards and threats can lead to similar consequences on the system [49-52] and, thus, be misclassified [44, 50, 53].

In the present Ph.D. work, the proposed modeling and simulation framework is aimed at specifically addressing the features (5) and (6) (safety and security, respectively), for which original and specific methodological solutions are needed to demonstrate the improved capabilities of CPSs on stand-alone physical systems.

1.2. Failures of the Cyber Physical Systems

CPSs must perform safely and securely [5, 47]. However, the CPSs can be subjected to failures, which translate in deviation from the expected behavior possibly leading to damages [54-56]. Failures may originate from both hazards and threats. Hazards are the presences of intrinsic property degradations, when occurred to CPSs components, can generate physical and cyber components failures and accidental events; threats are intentional actions that can impair both the physical and cyber parts of the system [18, 20]. Thus, CPSs failure analysis must comprise both safety and security aspects [18, 19, 44, 57].

1.2.1. Failures due to hazards

Hazards are the intrinsic properties or hazardous conditions that may cause harm or damage to systems, humans or environments [44, 58]. In regard to their origins, hazards of CPSs can be natural, anthropogenic or system operational events: natural hazards are the geological and meteorological phenomena, e.g., earthquakes, tsunami, that can suddenly disrupt the system functionality [59]; anthropogenic hazards relate to human errorous behaviors and activities leading the task outside its expectation [14, 60]; Aging and degradation under different operational conditions modify the way CPSs components work and interact with each other, generating multiple failure modes [61].

Failures of both hardware and software can compromise the functionality of CPSs. For example, sensors can degrade and fail in different modes such as bias, drift and freezing [61]; actuators can fail stuck, accidentally driving the physical process to be isolated from the controlling units of the cyber domain [62, 63].

Components failures can lead to two types of misoperations: (1) failure on-demand, e.g., failing to trigger protections or execute proper control strategies (when demanded); (2) malfunction, e.g., spurious triggering of protections (e.g., unintentional shutdown) or incorrect execution of control actions. Failures on-demand and malfunctions of both hardware and software components have gained increasing attention in the risk community [64, 65].

Resilience of CPS to failures can be granted by self-adaptiveness of control decisions on actuators, resorting to intelligent control systems that properly manipulate sensors measurements [66]. For example, Proportional-Integral-Derivative (PID) controllers, typically used as feedback controller in CPS to retroact to actuators the actions to be undertaken for responding to changes of physical parameters, may suffer of software failures/errors (generated from inadequate specification, incomplete testing scope and algorithm/logic failures) that are latent and triggered only when context modifications are to be met [64, 67]. In these situations, control rules adaptability to variable physical conditions is a fundamental requirement to the robustness of CPS for resilience during CPS operation.

1.2.2. Failures due to threats

Threats are intentional actions, i.e., attacks, to inflict harm or loss on the systems, humans or environments [44]. In CPSs, both physical (through local access) and cyber resources (through local or remote access) can be exploited in a threat action, to disrupt or destroy the computer equipment and the availability of data and to compromise the confidentiality, integrity and availability properties of CPSs [18, 68, 69].

CPSs reliance on digitalization and remote control systems increases their exposure to cyber attacks to controllers, databases, networks and human-system interfaces, that can result in the loss of system functionality. Malicious cyber attacks can be manifested as Denial of Service (DoS) [50, 70-72], False Data Injection (FDI) (e.g., packet/data modification) [73-75], network scan & sniffing [50, 76], integrity (e.g., through malware contagion) [77, 78] and illegal command executions [79]. They can be initiated in the cyber domain through local or remote accesses, mimicking the components failures but isolating the connectivity between cyber and physical systems, leaving the physical process uncontrolled and possibly drifting towards severe consequences.

Cyber attacks can cause serious security issues [80]. Under cyber attacks, e.g., by contagion of malware, security-related system features may result to be compromised and, the system safety potentially endangered. The identification of the cyber threats most affecting the system response is quite important for decision-making on optimal protection and resilience, as prevention and mitigation of malicious attacks contribute to guaranteeing CPS functionality [81-84].

1.2.3. Challenges in analyzing failures of the Cyber Physical Systems

In the context of CPSs, failure analysis faces to specific challenges [85]. For example, CPSs may be led to dangerous operational conditions not only due to unforeseen changes of the environmental conditions but also due to human errors, malicious attacks through the network and poor/wrong database use in support to parameters setting, thus, it is very important to achieve the robustness, resistance and adaptivity of the system to varying environmental conditions, and to some extend to mitigate against the vulnerabilities of cyber-attacks entering from real-time updating databases [12]; furthermore, while accidental hazards might be prognosed relying on the description of components behaviors for the estimates of failure times, remaining useful lifes, and failure events timeliness and sequencings [86, 87], cyber threats are relatively less accessible and less predictable due to the unknowns of the attackers plans and actions [18, 20].

In the present Ph.D. thesis, our main concern lies in the fact that stochastic hazards and malicious threats can lead to similar consequences on the system and, can be misclassified as component failures (and vice versa), disguising their character [49-52]. For example, in a situation where system shutdown is demanded, both failure of the shutdown of the actuator and interception of the shutdown command by an attacker result in unavailability of the safety action. In such situation, diagnosing the failure cause would allow taking the right decision to respond to the system shutdown unavailability with the right emergency procedure (e.g., manual operation of the actuation in the case of such cyber attack).

In fact, the failure analysis including both the safety and security aspects must address the following challenges:

I. Identification of hazards and threats most affecting the system functionality

Current treatment of the identification of hazards and threats focuses on the overall gathering of component (multiple) failure modes and vulnerabilities with the conditions simulating them to occur [88]. Efforts are mainly developed based on the analyst experience and brainstorming activity of the system analysis, through the use of expert judgment-based techniques, e.g., Failure

Mode Effects and Criticality Analysis (FMECA), hazard and operability (HAZOP), Bayesian Network (BN) [88, 89]. Besides this, Identification of the hazards and threats most affecting the system responses becomes more important in the anlysis of failures for CPSs, for understanding and identifying the conditions (represented by factors, parameters and variables values) that lead the system to critical conditions of failure, and preventing the accidents that may originate only if they are known in advance, at least to some extent [62, 78, 90-92].

II. Robustness in failure scenarios modeling and simulation

Modeling and simulation are used to explore and understand the behavior of a system, under different, possibly uncertain conditions, including hazardous and malicious ones [92, 93]. In system reliability assessment, binary-state graphic models (e.g., Markov Chain Model (MCM), Fault Tree Analysis (FTA) and Event Tree Analysis (ETA)) have been widely used for modeling components and system failures [94]. Whereas, a variety of conceptual or numerical models have focused on the formulation and modeling of malicious activities to CPSs, to understand the threats to the physical systems responses [18, 69, 80], e.g., graphical methods (such as attack graphs [95-97], attack trees [98], Petri nets [99]), mathematical models (such as those based on game theory [100, 101] and attacker-defender models [81, 83]), etc.

However, current treatment of failure scenarios modeling and simulations commonly neglects the impacts of physical degradation information in hazard analysis and, misses the attacker’s interests in injecting all possible failures to CPSs. Such problems have to be addresses in analyzing failures of the CPSs, to achieve the robustness of modeling and simulation.

III. Confidence in consequence analysis

Consequence analysis is aimed at evaluating and understanding the consequences of failures in CPSs, quantifying the estimates and/or the relevant ranges of the magnitudes of the consequences of failures, and determining the critical factors that most influence the consequences [88, 94, 102]. Conceptual, simulation or numerical methods have been developed to understand the physical phenomena that components failures and cyber threats lead to [88], but they hardly control the uncertainty affecting the system reliability (in hazard analysis) and responses to cyber threats [103].

In this sense, advanced methods are needed in consequence analysis to achieve a twofold potential benefit: on one hand, the more confident estimates for providing the analyst with the indication of what extent of damage the failures can lead to, and on the other hand, the allowance of balance between modeling efforts and computational demand with accuracy of the results.

IV. Effectiveness in protection design

CPSs protection design consists in the optimization of resource allocation for defensive barriers against (uncertain and unknown) cyber attacks [104-107] and the distinction bewteen cyber attacks and components stochastic failures [21, 102, 108, 109].

IV.1. Allocation of resources for defensive barriers against cyber attacks

A variety of defend-attack models have been proposed for this scope, focusing on the strategic interactions between defenders and attackers or/and the effectiveness of optimal defense resource allocations against adaptive cyber attacks. Graphical models (e.g., attack graphs [95, 110]) have been used to illustrate to a defender the proper security measures for defending the system. Potential system vulnerability paths that the attacker could exploit to gain access to a targeted cyber domain need to be identified and defended [95-97, 111-113]. Mathematical models (e.g., Copula-based models [82], a trilevel planner-attacker-defender model based on min-max-min optimization [81]) generally rely on a game-theoretical analysis and apply it to many areas (such as economics, political science, psychology, biology, computer science, and so on [114-116]), with the goal of advising the defender on the optimal allocation of defensive resources against attackers [80, 100, 117, 118].

However, all models mentioned above are developed from the viewpoint of a neutral opponent governing the attack/defense loss, under the strong assumptions of mutually consistent knowledge, rather than from the viewpoint of an intelligent adversary (attacker or defender) exploring the impacts of malicious (or self-interested) actions under uncertainty [104-107].

IV.2. Prompt recognition of cyber attacks from component stochastic failures

Cyber threats aimed at altering the CPS normal operation have been proposed to be diagnosed by scenario processing (i.e., modeling the malicious cyber events and their manifestation on the physical domain, affecting, in turn, both cyber and physical properties of the

CPS) [119-123]. A variety of methods for scenario processing specifically for diagnosing (rather than distinguishing component failures from) cyber attack have been proposed, based on artificial intelligence techniques. In general terms, observations are compared with the normal conditions measurements and a deviation from the legitimate data flow is found by methods such as the Sequential Probability Ratio Test (SPRT) [124, 125], the Cumulative Sum (CUSUM) chart [122, 126, 127], the Exponentially Weighted Moving Average (EWMA) inspection scheme [128], the Reversible-jump Markov Chain Monte Carlo (RJ-MCMC) [129], the control charts [130] and the transfer entropy [70].

Abovementioned methods can promptly recognize the predefined cyber attacks but, are not capable of distinguishing cyber attacks from component stochastic failures in CPSs. From the perspective of the analysis of failures in CPSs comprising both safety and security aspects, distinguishing cyber threats from component stochastic failures is important for anticipating the potential impact on the system functionality and defining proper protection and mitigation actions for resilience [21, 102, 108, 109]. To make CPSs resilient, the general modeling and simulation framework in the Ph.D. thesis is proposed to integrate the knowledge on safety, cyber security, defensive barriers and human interactions for addressing possible failures in a comprehensive and holistic way.

1.3. Research objectives of the thesis

The objective of the research activity presented in this Ph.D. thesis is to develop a general modeling and simulation framework for the analysis of failures of CPSs, considering both safety and security aspects, which includes: I. identification of hazards and threats (to identify the conditions that trigger anomalies in the systems and their causes), II. failure scenarios modeling and simulation (to characterize the system behavior under different operational conditions, including hazardous and malicious ones), III. consequence analysis (to explore the effects of stochastic component failures and cyber attacks onto the CPS functionality) and, IV. protection design (to take decisions on recovery measures for increasing system resilience).

I. Identification and prioritization of hazards and threats

Identification of hazards and threats of CPSs is performed for identifying the hazardous and malicious conditions that trigger anomalies in the systems and their causes. The identification of the most critical hazards and threats most affecting the system response is quite important for decision-making on optimal protection for preventing accidents that may originate; then, it is important to prioritize those CPSs components most affecting the system reliability and most vulnerable to cyber threats, to provide indication to the analyst of which CPS components deserve more attention.

II. Failure scenarios modeling and simulation

In the developed framework, we model and simulate failure scenarios to characterize the system behavior under different operational conditions and to account for the uncertainties affecting the system.

Integration of physical knowledge on aging and degradation into the modeling and simulation of CPSs components failures is expected to provide a better and more complete representation of the component degradation progression with respect to the traditional MCM method.

Similarly, modeling attacks aiming at damaging different components of the CPSs is exploring by simulation the generated different scenarios in the physical domain which lead to different consequences (e.g., magnitude of failure) are important to assess the CPS security with respect to cyber threats. Therefore, attack models are developed to launch attacks with different magnitudes and with the attackers' adaptive/responsive behaviors, to generate and explore specific deviations caused by cyber attacks on the CPS adapted as case study.

III. Consequence analysis

Consequence analysis evaluates the predicted outcome from an incident and the effect on its surrounding and people [131, 132]. It is used in risk management and assessment to understand the impact (e.g., range, magnitude, severity, etc.) of accidents and optimize system layout by improving design for reducing the risk from unacceptable levels.

In the present Ph.D. thesis, consequence analysis deals with the understanding of physical phenomena that the failure scenarios lead to and the exploration of their effects on the CPSs. To fully understand the consequences of failures on CPSs, on one side, we develop methods for quantifying and controlling the uncertainty affecting the system functionality and increasing confidence in consequence analysis, whereas on the other side, for exploring the most relevant hazards and threats affecting the CPS functionality that at the end are to be taken into account for decision-making and protection design.

IV. Protection design

Reducing the frequencies of occurrences of (disruptive) events, and of recovery measures for reducing the impacts of the accidents needs for the design of CPSs protection. On one side, defenders have to enforce defense strategies for allocation of resources for defensive barriers, against unknown and uncertain malicious cyber attacks [81, 82, 100, 117, 133], whereas, on the other side, since both stochastic failures and cyber attacks can compromise the CPS functionality, recognition of cyber attacks from component failures becomes paramount for increasing system protection and resilience [21, 81, 82, 102, 108].

1.4. Overview of the developed framework

Table 1 provides an overview of the research developed in the Ph.D. thesis. This will be introduced in the following paragraphs and, then, described in more details in the following Sections.

Table 1 Tasks and methods developed in the thesis

	Objectives	Hazard analysis		Threat analysis	
		Methods	Expectations	Methods	Expectations
I	Identification and prioritization of hazards and threats	**I.1** Sensitivity Analysis (SA)	• Identification of the components stochastic failures most affecting the CPS reliability	**I.2** MC-based exploration framework	• Generation of scenarios of cyber attacks to CPS components
II	Failure scenarios modeling and simulation	**II.1** Multi-State Physics Modeling (MSPM)	• Accurate component degradation modeling	**II.2** MC-based exploration framework	• Simulation of the effects of the cyber attacks on the system functionality
III	Consequence analysis	**III.1** Three-loop Monte Carlo (MC) simulation	• Exploration of the effects of components stochastic failures on the system reliability assessment	**III.2** Safety margins estimation approach	• Prioritization of the components most vulnerable to cyber attacks
IV	Protection design	**IV.1** Adversarial Risk Analysis (ARA) approach	• A novel prescriptive support strategy to optimize both the allocation of resources for defensive barriers against cyber attacks and the maintenance strategy to cope with component stochastic failures		
		IV.2 Non-Parametric Cumulative SUM (NP-CUSUM) approach	• Prompt distinction of cyber attacks from component failures in CPSs, for guiding decisions for the CPSs recovery from anomalous conditions		

1.4.1. Modeling and simulation for hazard analysis

Modeling and simulation for hazard analysis is here improved by proposing I.1 a Sensitivity Analysis (SA) for identifying the CPS components failures most affecting the system reliability in order to reduce the modeling efforts [134], II.1 a Multi-State Physics Modeling (MSPM) for accurate modeling relying on the integration of physical knowledge accounting for aging and degradation processes of the identified components [61], and III.1 a three-loop Monte Carlo (MC) simulation scheme for operationalizing the MSPM approach with respect to large scale CPSs and for quantifying and controling the confidence in reliability assessment, leveraging aging and degradation modeling with computational demand [103].

I.1. Sensitivity Analysis (SA)

To identify the components of the CPSs that most deserve accurate modeling accounting for aging and degradation process and for trading model accuracy and computational demand for reliability assessment, a SA is performed for the identification and prioritization of hazards.

SA can be performed in three different ways: local, regional and global [135, 136]. Global SA, in particular, measures the output uncertainty over the whole distributions of the input parameters and can be performed by parametric techniques, such as the variance decomposition method [137-140] and moment-independent method [141-143]. The variance-based method measures the part of the output variance that is attributed to the different inputs or set of inputs, without resorting to any assumption on the form of the model [135, 137, 144-146]. The moment-independent method, such as Hellinger distance and Kullback-Leibler divergence [147-149], allows quantifying the average effect of the input parameters on the reliability of the system and provides their importance ranking [150].

II.1. Multi-State Physics Modeling (MSPM)

To provides a better and more complete representation of the CPSs component degradation progression, specifically for that identified from the SA, MSPM is proposed for accurate component degradation modeling with realistic assumptions and available knowledge.

MSPM is a semi-Markov modeling framework that allows inserting physical knowledge on the system failure process, for improving the system reliability assessment by accounting for the effects of both the stochastic degradation process and the uncertain environmental and operational parameters [151-153]. In general, a MSPM describes the dynamics of component degradation in terms of transitions among a finite number M of degradation states, depending on a parameter vector δ. Similarly to Markov Chain Model (MCM), a state probability P is assigned to each degradation state, forming a state probability vector $P(t,\delta) = \{P_0(t,\delta), P_1(t,\delta), \cdots, P_j(t,\delta), \cdots, P_M(t,\delta)\}$ for all M states.

III.1. Three-loop Monte-Carlo (MC) simulation

To operationalize the MSPM approach, and to quantify and control the uncertainty affecting the system reliability model, a three-loop MC simulation scheme is proposed in consequence analysis for hazard analysis.

MC simulation [154] is a broad class of computational methodology for obtaining the estimates of the solution of risk analysis problems, e.g., failure time, system reliability, etc. Estimates of the MC simulation can be achieved by a large amount of repeated random samples sampled from the system state model.

In this work, the three-loop MC simulation scheme comprises three steps: (*i*) the identification of the components of the system for which a component-level MSPM is beneficial, because of the importance of the component for the system unreliability, (*ii*) the quantification of the uncertainties in the MSPM component models and their propagation onto the system-level model, and (*iii*) the selection of the most suitable modeling alternative that balances the computational demand for the system model solution and the robustness of the system reliability estimates.

1.4.2. Modeling and simulation for threat analysis

Modeling and simulation for threat analysis has been improved by proposing a MC-based exploration framework for I.2 generating and II.2 simulating the effects of cyber attack scenarios in CPSs, accounting for multiple failure modes of attacked components of the CPSs [84], and for III.2 prioritizing the components most vulnerable to cyber attacks by a safety margins estimation-based approach [84].

I.2. & II.2. The MC-based exploration approach

MC simulation allows considering the interactions among the physical parameters of the process (e.g., temperature, pressure, flow rate, etc.), human actions, components stochastic failures, and malicious activities [154]. Attacks aiming at damaging different components of the CPSs can, thus, be explored, generating different scenarios in the physical domain which lead to different consequences (e.g., magnitude of failure). Similarly, models can be introduced for describing attack magnitudes and the attackers' adaptive/responsive behaviors, generating and exploring specific deviations caused by cyber attacks.

A MC-based exploration framework can generate and process cyber attack scenarios in CPSs accounting for multiple failure modes of the attacked components, to test the effects of the cyber threats on the system functionality and integrity [84].

III.2. The safety margins estimation approach

A safety margin estimation approach of literature [155-157], that traditionally measures the minimum distance between the system loading and its capacity, can be undertaken for processing cyber attack scenarios, to estimate the extent of the consequences of cyber threats on the CPS components.

A number of non-parametric statistical methods have been used in safety analysis for safety margin estimation: the Wilk's method based on Order Statistics (OS) [158-161], Beran and Hall simple linear interpolation [162], Hutson fractional statistics [163] and data-based bootstrap method [164]. Among these, OS is popular and consolidated because it provides relatively conservative results with a few computer code runs, for leveraging the usually expensive computational cost of simulation codes [155, 159, 165]. In Task III.2, we take a "Bracketing" OS approach for tackling the computational problem and calculating the safety margins and to prioritize the most vulnerable components for cyber security protection decision-making [84].

1.4.3. Modeling and simulation for protection design

For design of CPSs protection, on one hand, an Adversarial Risk Analysis (ARA) approach is developed in IV.1 for obtaining a novel prescriptive defender support strategy that optimize both the allocation of resources for defensive barriers against cyber attacks and the maintenance strategy to cope with component stochastic failures [166]; and for online diagnostics of cyber attacks to CPSs, a Non-Parametric Cumulative Sum (NP-CUSUM) detection approach is developed in IV.2, to promptly recognize cyber attacks, distinguish them from component failures, and guiding decisions for CPSs recovery [53, 167].

IV.1. Adversarial Risk Analysis (ARA)

ARA builds on statistical risk analysis and game theory to analyze decision situations involving two or more intelligent opponents who make decisions under uncertainty [104, 106, 168, 169]. Different from the traditional game-theoretical models under the strong assumptions of mutually consistent knowledge between defender and attacker, ARA realistically assumes that each agent will only know his own beliefs and preferences and that these are not known to the others, and advises one player (e.g., defender) against the other(s) (e.g., attacker), within his own subjective expected utility model. ARA addresses this limitation by modeling and analyzing intelligent actors (attackers or defenders), for which the outcomes (or losses) in the game-theoretical model are uncertain [104, 106, 169].

In this thesis work, an ARA framework can provide a novel one-sided (i.e., defender) prescriptive support strategy for optimizing allocation of resources for the defensive barriers based on a subjective expected utility model.

IV.2. Non-Parametric Cumulative Sum (NP-CUSUM) approach

The NP-CUSUM approach is a sequential anomaly detection technique that allows for quick detection of parameter changes in physical systems [122, 170, 171], and has been proposed to embedded within the information systems for detecting faults [172, 173] and cyber attacks, e.g., DoS [122, 174], spamming [175], network scanning [176], etc.

In this thesis work, a consolidated NP-CUSUM approach is proposed for real-time diagnosing unknown cyber attacks and distinguishing them from stochastic failures of components of CPSs, by relying on the simultaneous treatment of the measurements taken from redundant channels.

To validate and actualize its capability in diagnosing cyber threats to CPSs, a reliability assessment is hereby performed for the NP-CUSUM-based cyber security diagnostic tool. The study takes simultaneously into account two fundamental aspects affecting the reliability assessment: (i) the uncertainty of the NP-CUSUM algorithm, and (ii) the modeling of the uncertainty related to the human operator cognition in interpreting and understanding the outcomes of the diagnostic tool. Human cognition will be modelled by Bayesian Belief Network (BBN) that structures the expert knowledge and understanding on the dependences among human factors (e.g., Performance Shaping Factors (PSFs)) and their causalities to the human cognition errors, in line with [177-182].

1.5. Case Studies

Without loss of generality and for demonstration purposes, the proposed modeling and simulation framework is demonstrated to nuclear CPSs (i.e., a typical Reactor Protection System (RPS) of NPPs, the digital I&C system of an Advanced Lead-cooled Fast Reactor European Demonstrator (ALFRED)). Subsections 1.5.1 and 1.5.2 are dedicated to a brief introduction of the considered case studies.

1.5.1. The Reactor Protection System

In NPPs, the RPS function is to trigger the NPP emergency shutdown, when an anomaly is detected in the measurements of a relevant signal (e.g., temperature). RPSs are identified to be the safety-critical systems and the embedded components deserve high priority of safety in NPPs [183]. In support to the implementation of risk-informed decision-making approaches, Probabilistic Safety Analysis (PSA) of modernizing NPPs demands for detailed dynamic models of digital I&C systems (e.g., RPSs) that can adequately represent digital components failure modes and quantify their contribution to the overall risk of the NPPs [64].

To this aim, more detailed modeling approaches are being increasingly integrated into existing PSA frameworks of RPSs safety assessment, for tackling the twofold purpose of PSA: on one side, the identification of the system failure domain and, on the other side, the quantification of the system failure probability.

In this thesis, we consider a typical RPS composed of two redundant channels (A and B) [184]. Each channel consists of one signal sensor, one Bistable Processor Logic (BPL) subsystem, and one Local Coincidence Logic (LCL) subsystem. Usually, redundancy is applied to sensors and signal processing units of RPS. However, with respect to the development of the modeling and simulation methods proposed in the thesis, we do not consider this for keeping the modeling complexity at a minimum without loss of generality.

Furthermore, the sensors S-A and S-B are considered to be Resistance Temperature Detectors (RTDs), because of the importance of these components in NPPs digital I&C systems [185, 186]. RTDs are safety-critical components and their effectiveness of detection of anomalous temperatures is very important for plant operators for monitoring the NPP operational conditions [187]. The reliability and accuracy of RTDs is important for controlling the NPP power rate with confidence, guaranteeing large power rates with sufficient safety margins [61, 186].

In this thesis, we consider the RPS as illustrative CPS case study in hazard analysis, to identify the embedded components that most deserve accurate modeling according to the SA, for accurate reliability assessment based on the MSPM. In Task III.1, the system-level MSPM of the RPS is operationalized in a three-loop MC simulation scheme, showing benefits of integrating the physics knowledge into the system reliability modeling.

1.5.2. The Advanced Lead-cooled Fast Reactor European Demonstrator and its digital Instrumentation and Control system

ALFRED is a small-size (300 MW) pool-type lead-cooled fast reactor, cooled by molten lead to ensure the favourable physical features and realize a simplified plant layout [188]. At full power nominal conditions, the dynamics processing of the ALFRED primary and secondary cooling systems is controlled by a multi-loop PI (Proportional and Integral) control scheme, i.e., a decentralized control scheme, because of its simplicity of implementation and robustness to malfunctioning of the single control loops [189, 190]. Both feedback and feedforward digital control schemes are adopted for ALFRED. The PI-based feedback control configuration employs four SISO (Single Input Single Output) control loops independent of each other [191].

Both stochastic components failures and cyber attacks can compromise the correct functionality of the CPSs. Cyber attacks manifest themselves in the physical system and, can be misclassified as component failures, leading to wrong control actions and system responses. In this Ph.D. work, without loss of generality, we consider the possible occurrence of stochastic components failures and cyber attacks in the digital I&C system of the ALFRED, whose previously developed object-oriented DYMOLA simulator [189, 190] with the control scheme is utilized for simulating the ALFRED dynamic response to failures and cyber attacks.

The safety margin estimates of the cyber breach samples in threat analysis allows exploring the system behaviors under different cyber attacks and identifying the most vulnerable components of the ALFRED digital I&C system. In protection design, the ARA can advise the defender of the ALFRED the optimal portfolio of allocation of resources for defensive barriers, minimizing the system integrity loss against uncertain cyber attacks, on one hand, and on the other hand, the NP-CUSUM online diagnostic tool allows for promptly recognizing cyber attacks from component failures in ALFRED, and guiding decisions for the ALFRED recovery from anomalous conditions.

1.6. Thesis Structure

Figure 2 shows the structure of the thesis work. Chapters 2, 3 and 4 are dedicated to the research objectives introduced in Section 1.4 and, Chapter 5 draws the conclusions and future perspectives. At the end, a collection of the international peer-reviewed journals papers finalized during the Ph.D. is included for further details.

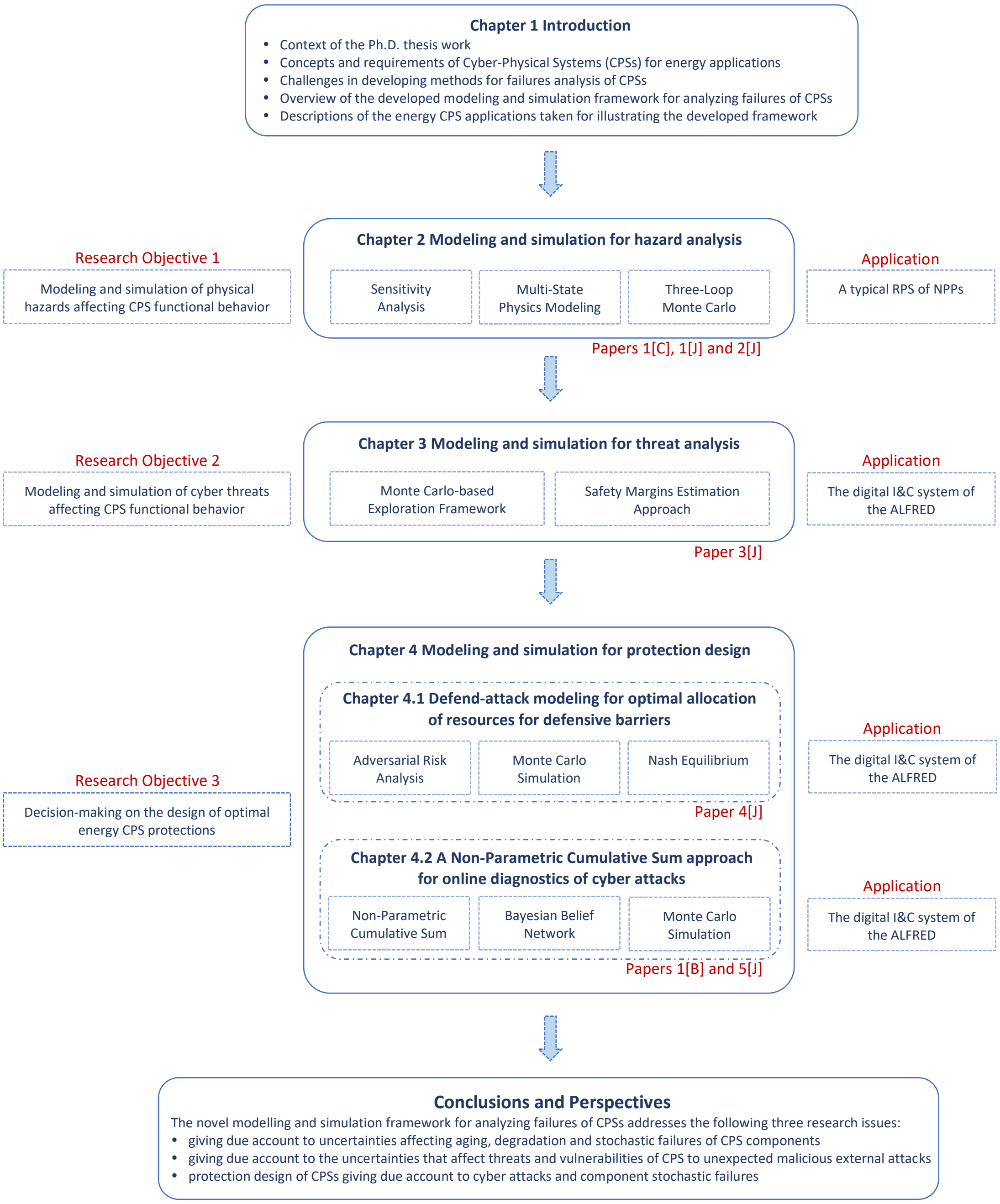

Chapter 1 Introduction
Context of the Ph.D. thesis work
Concepts and requirements of Cyber-Physical Systems (CPSs) for energy applications
Challenges in developing methods for failures analysis of CPSs
Overview of the developed modeling and simulation framework for analyzing failures of CPSs
Descriptions of the energy CPS applications taken for illustrating the developed framework
Research Objective 1
Modeling and simulation of physical hazards affecting CPS functional behavior
Chapter 2 Modeling and simulation for hazard analysis
Sensitivity Analysis
Multi-State Physics Modeling
Three-Loop Monte Carlo
Application
A typical RPS of NPPs
Papers 1[C], 1[J] and 2[J]
Research Objective 2
Modeling and simulation of cyber threats affecting CPS functional behavior
Chapter 3 Modeling and simulation for threat analysis
Monte Carlo-based Exploration Framework
Safety Margins Estimation Approach
Application
The digital I&C system of the ALFRED
Paper 3[J]
Chapter 4 Modeling and simulation for protection design
Chapter 4.1 Defend-attack modeling for optimal allocation of resources for defensive barriers
Adversarial Risk Analysis
Monte Carlo Simulation
Nash Equilibrium
Application
The digital I&C system of the ALFRED
Paper 4[J]
Research Objective 3
Decision-making on the design of optimal energy CPS protections
Chapter 4.2 A Non-Parametric Cumulative Sum approach for online diagnostics of cyber attacks
Non-Parametric Cumulative Sum
Bayesian Belief Network
Monte Carlo Simulation
Application
The digital I&C system of the ALFRED
Papers 1[B] and 5[J]
Conclusions and Perspectives
The novel modelling and simulation framework for analyzing failures of CPSs addresses the following three research issues:
giving due account to uncertainties affecting aging, degradation and stochastic failures of CPS components
giving due account to the uncertainties that affect threats and vulnerabilities of CPS to unexpected malicious external attacks
protection design of CPSs giving due account to cyber attacks and component stochastic failures

Figure 2 Sketch of the thesis structure

SECTION II. DETAILS OF THE DEVELOPED FRAMEWORK

This Section consists in 3 Chapters (i.e., Chapters 2 Modeling and simulation for hazard analysis, Chapter 3 Modeling and simulation for threat analysis, Chapter 4 Modeling and simulation for protection design, and Chapter 5 Conclusions and future perspectives) that describe in details the original contributions resulting from the Ph.D. research work.

2. MODELING AND SIMULATION FOR HAZARD ANALYSIS

Contents of the Chapter have been adapted from:

1[C] *Wang, W., Di Maio, F. and Zio, E., 2017. A sensitivity analysis for the adequacy assessment of a multi-state physics modeling approach for reliability analysis. In 26th European Safety and Reliability Conference, ESREL 2016 (pp. 465-472). CRC Press/Balkema.*

1[J] *Wang, W., Di Maio, F. and Zio, E., 2016. Component-and system-level degradation modeling of digital Instrumentation and Control systems based on a Multi-State Physics Modeling Approach. Annals of Nuclear Energy, 95, pp.135-147.*

2[J] *Wang, W., Di Maio, F. and Zio, E., 2017. Three-loop Monte Carlo simulation approach to Multi-State Physics Modeling for system reliability assessment. Reliability Engineering & System Safety, 167, pp.276-289.*

In hazard analysis, we develop methods to give due account to uncertainties affecting aging, degradation and stochastic failures of CPS components. A Sensitivity Analysis (SA) approach is performed for identifying the component stochastic failures most affecting the CPS reliability in Section 2.2, which deserve accurate modeling of aging and degradation by a Multi-State Physics Modeling (MSPM) approach in Section 2.3. Then, in Section 2.4, a three-loop Monte Carlo (MC) simulation scheme is developed to explore the effects of component stochastic failures on the system reliability assessment, leveraging aging and degradation modeling with computational demand.

2.1. Case study: the Nuclear Power Plant (NPP) Reactor Protection System (RPS)

The RPS of a NPP, described in Section 1.5.1, is considered as case study for numerical evaluation. As shown in Figure 3, the RPS is composed of two redundant channels (A and B). Each channel consists of one signal sensor (S-A and S-B), one BPL subsystem (BPL-A and BPL-B), and one LCL subsystem (LCL-A and LCL-B).

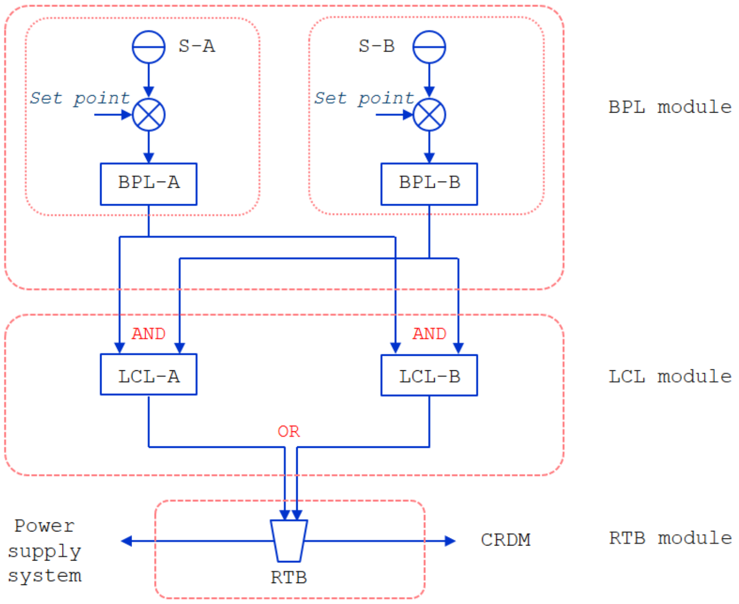


Figure 3 RPS scheme [184]

If any one of the two redundant measured signals exceeds a triggering threshold value, a Partial Tripping Signal (PTS) is sent to the corresponding BPL. The signal processing activates only if both channels produce the PTS: each PTS from a BPL is sent to both LCL-A and LCL-B, which process information by an "AND" gate. In other words, an Emergency Shutdown Signal (ESS) is produced only when receiving two PTSs from different BPLs; ESSs, then, activate the Reactor Trip Breaker (RTB), when at least one ESS is triggered, i.e., the information is processed by an "OR" gate. Once the RTB is activated, the power supply system and Control Rod Drive Mechanism (CRDM) which are connected with the RTB activate to control the power of the reactor.

According to the RPS scheme of Figure 3, three modules are identified:

- The BPL Module consists of two groups of components: sensor and BPL (i.e., "S-A and BPL-A" and "S-B and BPL-B"); these components are connected in series and their failure effects on the system can be combined.
- The LCL Module consists of the two LCLs (i.e., LCL-A and LCL-B); since the ESS is triggered only when both LCLs simultaneously receive two PTSs from the two BPLs, this module is highly dependent of the BPL module.
- The RTB Module.

2.1.1. The traditional RPS Markov Chain Model

In this Section, a binary-state MCM is built as reference for the reliability assessment of the RPS. To do this, intra- and inter-module states leading to the system failure are identified. Intra-module states refer to events leading to the system failure that concerns components belonging to the same module; inter-module states relate to system failures from combined component events in different modules.

Figure 4 shows the RPS-MCM, whose states (listed in Table 2) are grouped into four categories that relate to the intra- and inter-module distinction. The following assumptions have been made for the subsequent quantitative analysis:

- Transitions can occur from the system functioning state (state 0) to any of the absorbing failure states of the intra-module category and from the intermediate state (state 3) to any of the absorbing states of the inter-module category. The transition rates are taken from public databases [192, 193] and reported in Table 3.
- No repairs are considered.

Table 2 Component states

State	Description
0	RPS functioning state.
1	Either one of the RTD sensors fails.
2	Either one of the BPLs fails to send out PTSs.
3	Either one of the LCLs fails to produce the ESS.
4	RTB fails.
5	One LCL has failed and, then, one sensor fails.
6	One LCL has failed and, then, one BPL fails.
7	Both LCLs fail to produce the ESS.
8	One LCL has failed and, then, the RTB fails.
9	Common cause failure of BPL-A and BPL-B.
10	Common cause failure of LCL-A and LCL-B.

Table 3 Transition rates

Symbol	Description	Value (/yr)
λ_S	RTD failure rate	8.760e-1 [192]
λ_B	BPL failure rate	8.760e-3 [192]
λ_L	LCL failure rate	4.380e-2 [192]
λ_R	RTB failure rate	3.767e-4 [193]
β	Common cause factor	0.1
λ_{BS}	BPL self-fault failure rate	$(1-\beta)*\lambda_B$=7.884e-3
λ_{LS}	LCL self-fault failure rate	$(1-\beta)*\lambda_L$=3.942e-2
λ_{BC}	BPLs common cause failure rate	$\beta*\lambda_B$=8.760e-4
λ_{LC}	LCLs common cause failure rate	$\beta*\lambda_L$=4.380e-3

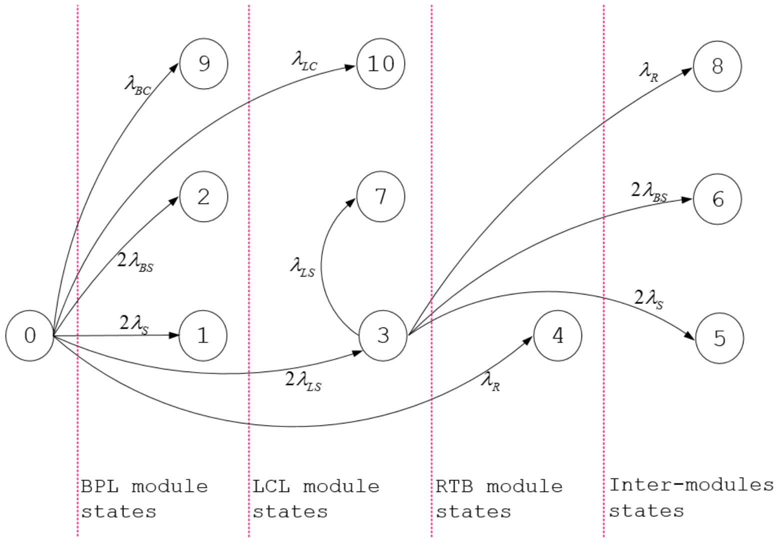


Figure 4 The RPS-MCM where states are grouped according to their intra-module and inter-modules characteristics

The RPS unreliability $P(t)$, and the individual modules unreliabilities $P_{BPL}(t)$, $P_{LCL}(t)$, $P_{RTB}(t)$ and $P_{Inter\text{-}modules}(t)$ are presented in Figure 5. A visual analysis of the unreliability curves shows that most of the system unreliability $P(t)$ is contributed by the BPL, that is to say, the absorbing states of the BPL module most contribute to the system unreliability.

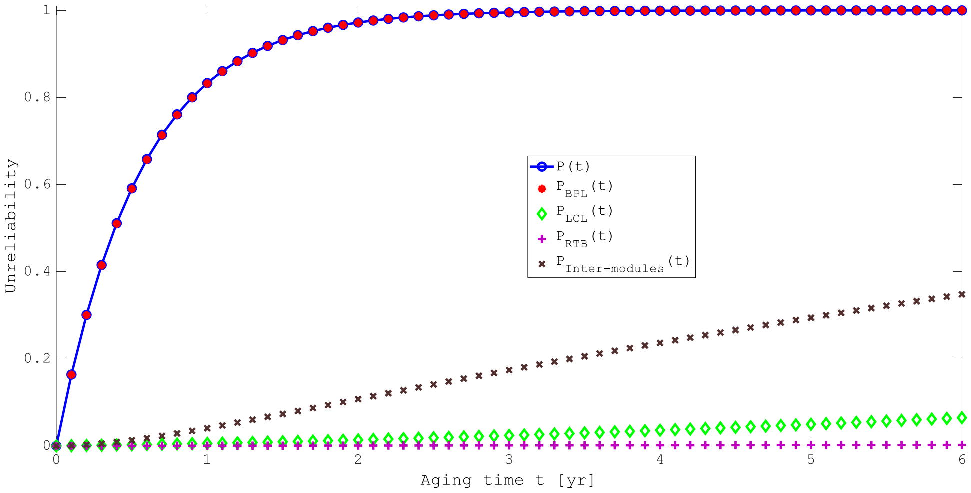


Figure 5 Unreliability curves of RPS and its modules

2.1.2. Uncertainty assessment

The standard deviation values of the transition rates of Table 3 are either provided by public databases or can be estimated by resorting to Fisher Information [194, 195]. The procedure for this is here described with reference to the RTD, whose failure rate standard deviation is not provided in [192]:

- Simulation of life tests.

Notice that the Cumulative Distribution Function (CDF) describing the uncertain timing of the RTD New-to-drift failure mode reaches its failure time (i.e., the unreliability value turns out to be equal to 1) at 5.8yr [61]. Therefore, in this work, we take a mission time *T*=6yr as the end of the right-censored life tests. We randomly sample N_R=1000 trials of RTD failure times from an exponential distribution with constant transition rate λ_S (Table 3). If the sampled time exceeds the mission time *T*=6yr, the test is considered right-censored [94].

- Estimation of the standard deviation $\hat{\sigma}_s$ of λ_S.

The variance of λ_S can be estimated based on the observed Fisher information. The Fisher Information Matrix is defined from the Maximum Likelihood function or its LogLikelihood, and can be estimated by [94]:

$$\log L\left(t,\hat{\lambda}_s\right)=\log\left(\prod_i f_T\left(t_i;\hat{\lambda}_s\right)\cdot\prod_j R\left(t_j;\hat{\lambda}_s\right)\right) \qquad (2\text{-}1)$$

where *i* and *j* are the RTD failure times before *T* and the times right-censored by *T*, respectively, and $f_T\left(t_i;\hat{\lambda}_s\right)$ and $R\left(t_j;\hat{\lambda}_s\right)$ are the RTD failure time probability density function (pdf) and the RTD reliability:

$$f_T\left(t_i;\hat{\lambda}_S\right)=\hat{\lambda}_S\cdot e^{-\hat{\lambda}_S t_i} \qquad (2\text{-}2)$$

$$R\left(t_j;\hat{\lambda}_S\right)=e^{-\hat{\lambda}_S t_i} \qquad (2\text{-}3)$$

With respect to the observable random failure time *t*, the Fisher Information Matrix $J\left(\hat{\lambda}_s\right)$ can be expressed as:

$$J\left(\hat{\lambda}_S\right)=E\left[\left(\frac{\partial\log L\left(t;\hat{\lambda}_S\right)}{\partial\hat{\lambda}_S}\right)^2\right] \qquad (2\text{-}4)$$

As a result, the variances of the parameters $\hat{\lambda}_s$ can be provided from the main diagonal of its inverse matrix $J^{-1}(\hat{\lambda}_s)$, namely, the estimated standard deviations $\hat{\sigma}_s$ of the parameters:

$$\hat{\sigma}_S = J^{-1}\left(\hat{\lambda}_S\right) \tag{2-5}$$

Under the condition of mild regularity, $J^{-1}(\hat{\lambda}_s)$ can be calculated by Eq.(2-6):

$$J^{-1}\left(\hat{\lambda}_S\right) = \left[-E\left(\frac{\partial^2 \log L\left(t;\hat{\lambda}_S\right)}{\partial \hat{\lambda}_S{}^2}\right)\right]^{-1} \tag{2-6}$$

and the standard deviation can be estimated as:

$$\hat{\sigma}_S = J^{-1}\left(\hat{\lambda}_S\right) = \left[-E\left(\frac{\partial^2 \log L\left(t,\hat{\lambda}_S\right)}{\partial \hat{\lambda}_S{}^2}\right)\right]^{-1} \tag{2-7}$$

The standard deviations of the transition rates of the BPLs, LCLs, and RTB are also estimated by the Fisher Information Methodology (Table 4).

Table 4 Estimated transition rates

Symbol	Mean value (/yr)	Standard deviation (/yr)
λ_S	8.760e-1	7.720e-1
λ_B	8.760e-3	7.867e-8
λ_L	4.380e-2	1.981e-6
λ_R	3.767e-4	1.332e-10

2.1.3. Uncertainty propagation

Uncertainty in binary transition rates is propagated through the RPS-MCM as follows:

1) Set initial time t_0=0 and mission time T=6yr, and partition the time axis into small intervals of length dt=0.01yr;

2) Sample the component failure rates from the Gaussian distributions $N\left(\lambda_k, \hat{\sigma}_k\right)$ that are shown in Table 3, where, $k = S, B, L, R$;

3) For each time instant t before T, compute the system unreliability from the MCM [196];

$$P\left(t\mid\lambda_S,\lambda_B,\lambda_L,\lambda_R\right)=1\text{-}\left(1+\frac{2\left(1\text{-}\beta\right)\lambda_L\left(e^{\left(\beta\lambda_B+\lambda_L\right)t}\text{-}1\right)}{\left(\beta\lambda_B+\lambda_L\right)}\right)e^{\text{-}\left(2\lambda_S+\left(2\text{-}\beta\right)\lambda_B+\left(2\text{-}\beta\right)\lambda_L+\lambda_R\right)t} \tag{2-8}$$

4) Repeat the steps 2) and 3) for *Na*=1000 times;

5) Compute the 5th and 95th percentiles for each time instant *t*.

Figure 6 shows the plot of the pointwise double-sided 90% confidence interval of the system unreliability. The confidence interval is large all over the system life *T*, because of the large uncertainty that affects the MCM transition rates due to the weak knowledge utilized to build the, therefore, quite inaccurate RPS-MCM.

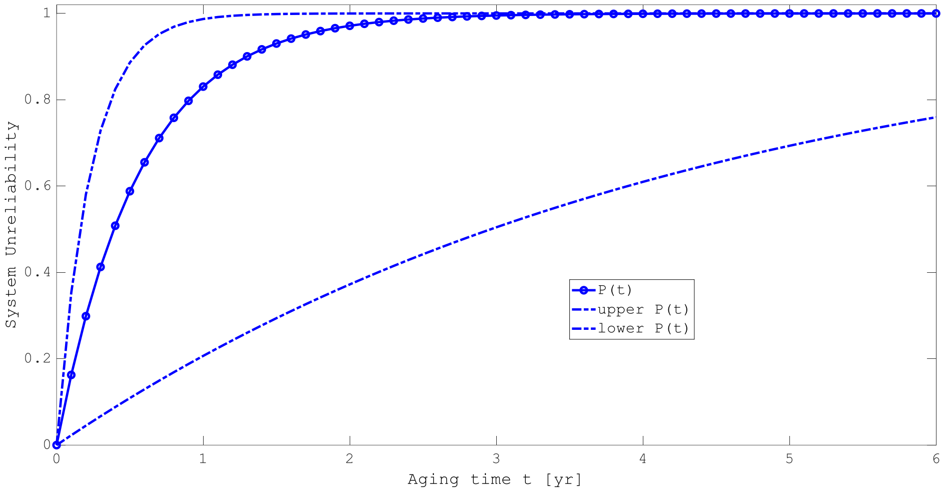


Figure 6 Confidence intervals from the RPS-MCM system unreliability

2.2. Sensitivity analysis

The purpose of the SA in this work is to identify the components of the RPS that most deserve accurate modeling of aging- and degradation-dependent transition rates, for accurate system reliability assessment [134] and for trading model accuracy and computational demand for practical reliability assessment.

With reference to the RPS of Figure 3, we describe the SA as follows:

1) Calculate the moment-independent sensitivity measures between the unreliability *P*(*t*) of the RPS and the unreliability $P_k(t)$ of its *k*-th module contributor (i.e., $P_{BPL}(t)$, $P_{LCL}(t)$, $P_{RTB}(t)$ and $P_{Inter\text{-}modules}(t)$ (noted that inter-modules refers to the system states that affect

simultaneously components of different modules)), to identify the most important module in the system;

2) Calculate the moment-independent measure for the sensitivity between the module unreliability $P_k(t)$ and the unreliability of its l-th embedded component $P_l(t)$, to identify the component most affecting the module unreliability.

The SA is here adopted based on moment-independent sensitivity measures [197], such as Hellinger distance and Kullback-Leibler divergence [147-149], which rest on the common rationale that the sensitivity measures can be computed as expected generalized distances between the output distribution and the conditional output distribution given the model input(s) of interest [198]. In detail, the Hellinger distance $H_k[p(t),p_k(t)]$ measures the difference between the pdf $p(t)$ of the system unreliability and the pdf $p_k(t)$ of the k-th contributor to the system failure, i.e., BPL, LCL, RTB, Inter-modules [148, 149]:

$$H_k\left[p(t),p_k(t)\right]=\left[\frac{1}{2}\int\left(\sqrt{p(t)}-\sqrt{p_k(t)}\right)^2 dt\right]^{\frac{1}{2}}=\left[1-\int\left(\sqrt{p(t)\cdot p_k(t)}\right)^2 dt\right]^{\frac{1}{2}} \tag{2-9}$$

The k-th contributor is important if H_k is small.

The Kullback-Leibler divergence $KL_k[p(t),p_k(t)]$ measures the different information carried by the pdf $p(t)$ of the system failure and the pdf $p_k(t)$ of the k-th contributor according to Eq. (2-10) [148, 149]:

$$KL_k\left[p(t),p_k(t)\right]=\int_{-\infty}^{+\infty}p(t)\log\left(\frac{p(t)}{p_k(t)}\right)dt \tag{2-10}$$

with the values in $[0,+\infty]$. In practical cases, the symmetric form of Kullback-Leibler divergence can be untilized as follows [199]:

$$KL_{sym,k}\left[p(t),p_k(t)\right]=KL_{sym,k}\left[p_k(t),p(t)\right]=\frac{1}{2}KL_k\left[p(t),p_k(t)\right]+\frac{1}{2}KL_k\left[p_k(t),p(t)\right] \tag{2-11}$$

The k-th contributor is important if $KL_{sym,k}$ is small, in relative terms.

2.2.1. Sensitivity analysis results

Table 5 lists the Hellinger distance and Kullback-Leibler divergence values for each module contributor to the system unreliability, respectively: both measures identify the BPL as the most important contributor.

Table 5 Ranking of contributors to the RPS unreliability

Input	H_k	$KL_{sym,k}$
Intra-BPL	0.0013	6.4539e-6
Intra-LCL	0.6398	2.4181
Intra-RTB	0.6872	3.7300
Inter-Module	0.6000	1.8809

Since the BPL module plays the most significant role in affecting the reliability of the RPS, we now focus on identifying the BPL component most contributing to its failure. To rank the importance of the l-th component embedded in the BPL module, the two SA measures of Eqs. (2-9) and (2-11) are quantified. The sensors turn out to be the most important components contributing to the BPL module unreliability (see Table 6).

Table 6 Ranking of the contributors to the BPL unreliability

Input	H_l	$KL_{sym,l}$
Sensors	0.2391	0.2460
BPLs	0.6219	2.1599

2.3. The Resistance Temperature Detector (RTD) Multi-State Physics Modeling (MSPM)

The results of the SA point at the sensor (i.e., the RTD) as the component deserving more modeling efforts for accurate RPS unreliability estimation. A component MSPM is here developed to describe the RTD degradation-to-failure process, inserting physics knowledge in the model.

As discussed in [61, 200], among the RTDs failure modes (e.g., bias, drift, performance degradation, freezing and calibration error), experimental evidence suggests that the main failure mode is drift. Drift is measured by the response time τ that the RTD needs to reach 63.2% of a sudden temperature change of the RTD. Aging t and air gap size δ between the bottom of the thermowell and the sensing tip (that changes because of contamination and mechanical shocks) are the most likely contributors to the drift [187]. The response time $\tau(t,\delta)$ is assumed not to exceed the RTD failure threshold γ_Y during normal operation and in relation to this, the RTD failure boundary is defined as $\partial F = G(t,\delta) = 0$, where,

$$G(t,\delta) = \tau(t,\delta) - \gamma_Y \tag{2-12}$$

The RTD-MSPM shown in Figure 7 depicts, in a two-state diagram, the partition by ∂F of the safe domain S from the failure domain F of the RTD. The RTD-MSPM assumptions are described as follows:

- S_0^{RTD} is the RTD functioning state and S_1^{RTD} is the RTD drift failure state;
- Transitions can occur between the two states with failure rate $\lambda_S(t \mid \delta)$ and repair rate $\mu_S(t \mid \delta)$, functions of the time t and the affecting factor δ;
- At the initial time t=0, the RTD is in its initial functioning state S_0^{RTD}.

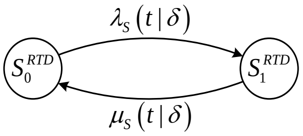


Figure 7 The RTD-MSPM model

To estimate the aging- and degradation-dependent transition rate $\lambda_S(t|\delta)$, we build the empirical relationship between τ, t and δ, i.e., $\tau(t,0)$ and $\tau(0,\delta)$, based on experimental data listed in Tables 7 and 8 [186, 187].

Table 7 Experimental data for τ at fixed t and δ=0 [186]

Aging Time t [yr]	0	2	4	5	6
Mean Response Time τ [s]	2.1	4.4	4.8	5.0	5.2
Standard Deviation $\sigma(t,0)$	1.67	0.77	0.72	0.77	0.67

Table 8 Fitted τ at t=0 and discrete δ based on experimental data from [186, 187]

Air gap size δ [mm]	0	0.4	0.8	1.2	1.6	2.0	2.4
Mean Response time t [s]	2.10	3.80	4.97	5.93	7.02	8.58	10.95
Standard Deviation $\sigma(0,\delta)$	1.18	1.19	1.64	2.47	3.61	4.98	6.51

An analytical function of $\tau(t,\delta)$ can be obtained relying on [61]:

$$\overline{\tau}\left(t,\delta\right)=\alpha_{t}\cdot\overline{\tau}\left(t-1,\delta\right) \tag{2-13}$$

$$\sigma^{2}\left(t,\delta\right)=\sigma^{2}\left(t-1,\delta\right)+\alpha_{t}^{2} \tag{2-14}$$

where $\overline{\tau}\left(t,\delta\right)$ is the mean value of the response time of Table 8, $\sigma(t,\delta)$ is its standard deviation, and the factor α_t accounts for the changes of response time τ with the increase of t, by scaling the $\overline{\tau}\left(t,0\right)$ using the scale factor α_t:

$$\alpha_{t}=\frac{\overline{\tau}\left(t,0\right)}{\overline{\tau}\left(t-1,0\right)} \tag{2-15}$$

where, $\alpha_{1}=\overline{\tau}\left(1,0\right)/\overline{\tau}\left(0,0\right)$.

The function $\tau(t,\delta)$ consists in a surface fitted to realizations of $\tau(t,\delta)$ sampled from the assumed Gaussian distributions with mean values $\overline{\tau}\left(t,\delta\right)$ and standard deviations $\sigma(t,\delta)$ at each discrete point, as shown in Figure 8 where one trial surface is plotted.

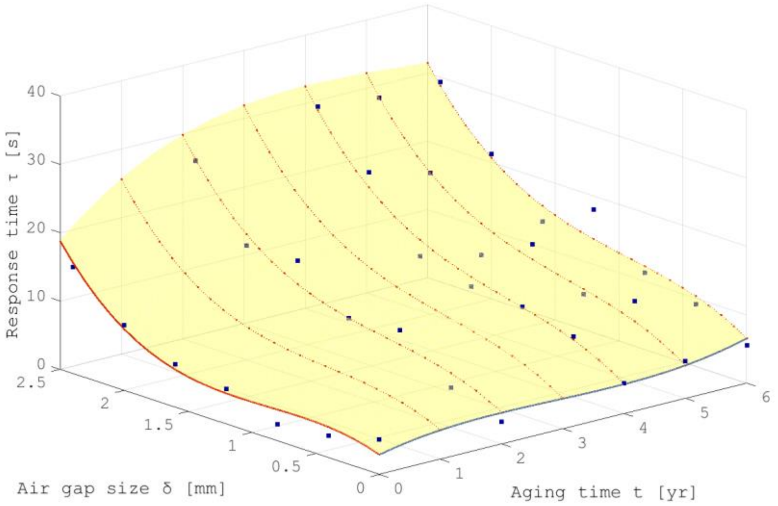


Figure 8 Fitted surface of τ(t,δ)

As mentioned in [187], the τ of a well-type RTD usually ranges in [4s, 8s]; hence, the RTD failure threshold γ_Y is here set equal to 8s. The Cumulative Distribution Function (CDF) $P_S(t, \delta)$ of the RTD new-to-drift-failure mode that can account for the stochasticity of the process and of the uncertainties affecting the degradation (for example, the initial air gap size δ_0 and the noise affecting the air gap size δ_t due to the vibration) can be found by running N_b(=1e4) Monte Carlo simulations, as follows:

- For each trial, at the initial time t=0, we sample the value of δ_0 from the uniform distribution $U(0,1)$ as initial air gap size.
- At each t that increases with the time step dt = 0.01yr, the value $d\delta_t$ is sampled from a normal distribution $N(0,0.025t)$; thus, $\delta = \delta_0 + d\delta_t$.
- At each t within the mission time $[t_0,t_m]$ (t_m = 6yr), τ is estimated using the curve $\tau(t, \delta)$ of Figure 8. If the value of τ exceeds the threshold γ_Y, the RTD is assumed to fail at time t with air gap size δ.

Pictorially, we can show the evolution of $\tau(t,\delta)$ for each trial, as sketched in Figure 9: for a sampled δ_0 (equal to 0.12mm), the air gap size oscillates during the RTD life around δ_0 (see in Figure 9(a)); on the other hand, the response time τ, stochastically changes with the increase of time t (in Figure 9(b)) and, thus, the transition between states S_0^{RTD} and S_1^{RTD} of drift failure mode is determined when the response time τ reaches the failure threshold γ_Y, as shown in Figure 9(c).

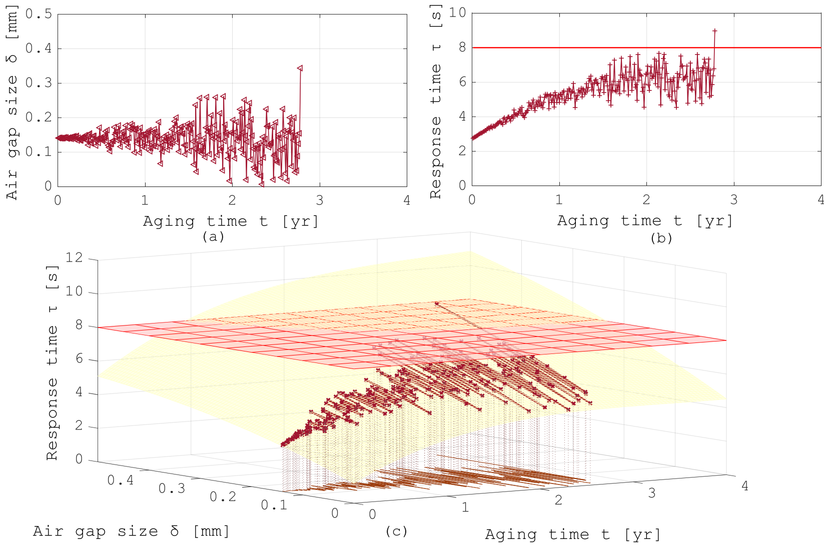


Figure 9 One trial of MC simulation: (a) the stochastic path of air gap size δ changing with the aging t; (b) the evolution of response time τ with the aging t; (c) the simulated path response time τ with respect to δ and t on the safety domain of the fitting curved surface

After the N_M trials of MC simulations have been run, the conditional PDF $p_S(t|\delta)$ and conditional CDF $P_S(t|\delta)$ of the RTD New-to-drift transition of Figure 7 can be empirically built (shown in Fig. 10 and 11, respectively) and used to calculate the conditional failure rate $\lambda_S(t|\delta)$:

$$F_S\left(t<T<t+\Delta t\,|\left(T>t,\delta\right)\right)=\frac{F_S\left(t<T<t+\Delta t\,|\,\delta\right)}{F_S\left(T>t\,|\,\delta\right)}=\frac{F_S\left(T<t+\Delta t\,|\,\delta\right)-F_S\left(T<t\,|\,\delta\right)}{1-F_S\left(T<t\,|\,\delta\right)}=\frac{P_S\left(t+\Delta t\,|\,\delta\right)-P_S\left(t\,|\,\delta\right)}{1-P_S\left(t\,|\,\delta\right)} \tag{2-16}$$

$$\lambda_S\left(t\,|\,\delta\right)=\lim_{\Delta t\to 0}\frac{F_S\left(t<T<t+\Delta t\,|\left(T>t,\delta\right)\right)}{\Delta t}=\lim_{\Delta t\to 0}\frac{P_S\left(t+\Delta t\,|\,\delta\right)-P_S\left(t\,|\,\delta\right)}{\Delta t}\cdot\frac{1}{1-P_S\left(t\,|\,\delta\right)}=\frac{p_S\left(t\,|\,\delta\right)}{R_S\left(t\,|\,\delta\right)} \tag{2-17}$$

For the sake of clarity, the conditional Probability Density Function (PDF) $p_S(t|\delta)$ and conditional Cumulative Distribution Function (CDF) $P_S(t|\delta)$for the surface of Figure 9, obtained by simulating Nb=1000 different degradation processes, are plotted in Figures 10 and 11.

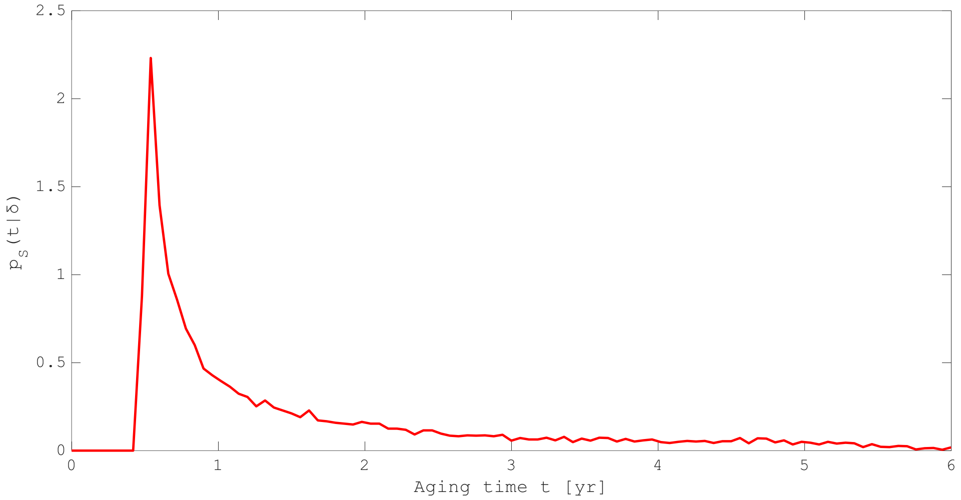


Figure 10 Conditional probability density function of RTD New-to-drift failure mode

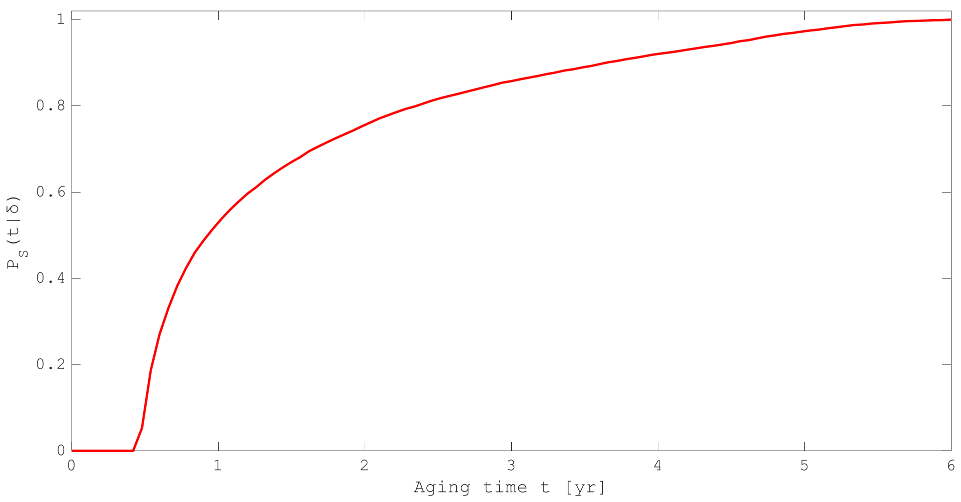


Figure 11Conditional cumulative distribution function of RTD New-to-drift failure mode

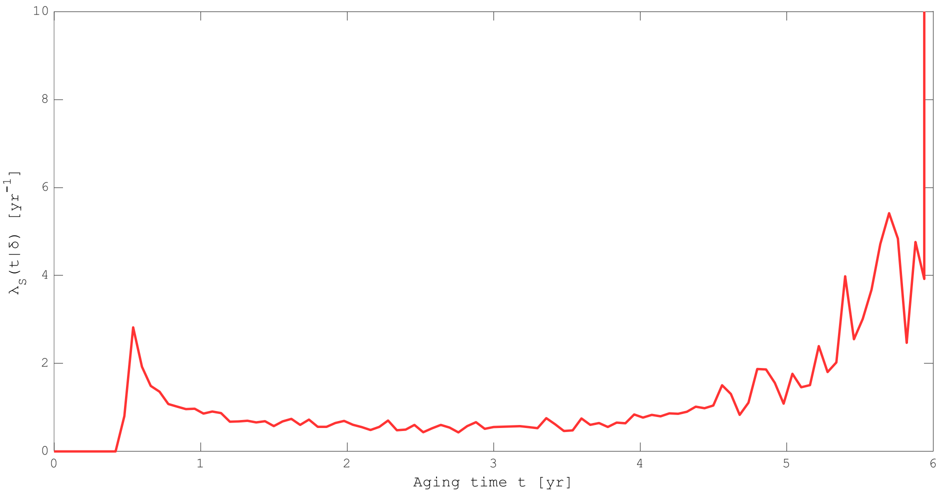


Figure 12 Conditional failure rate $\lambda_S(t|\delta)$ of RTD new-to-drift failure mode

It is worth mentioning that the conditional CDF $P_S(t|\delta)$ of Fig. 11 describing the uncertain timing of RTD New-to-drift failure mode shows a sharp increase in [0.5, 1.5] yr, after which it starts to level off to reach $P_S(t|\delta)$ at 5.8yr. Therefore, the failure rate $\lambda_S(t|\delta)$ of Fig. 12 shows the typical infant mortality and wear out periods, and tends to be constant in the useful life, which coincides with a general bath-tub curve, but with non-constant values along life.

2.4. Three-loop Monte Carlo simulation

The RPS-MSPM model of Figure 13 embeds the RTD-MSPM model of Figure 7, while components other than the RTD are assumed to obey binary-state behaviors as in the reference MCM of Figure 4.

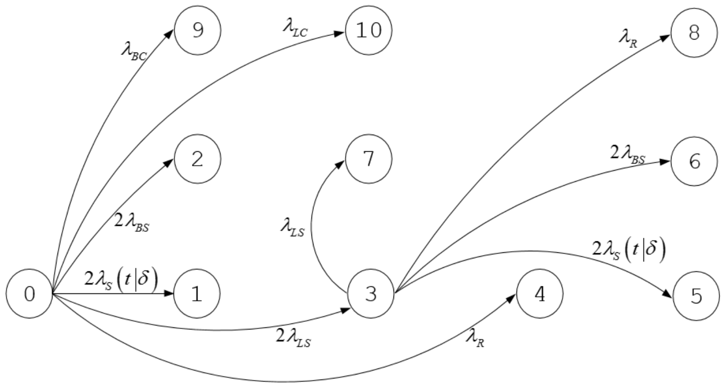


Figure 13 The RPS-MSPM integrating the RTD-MSPM

We propose the three-loop MC simulation for the RPS reliability assessment, with confidence quantification related to the uncertainty in the RTD physical parameters propagated through the surfaces $\tau(t,\delta)$ of Section 2.3, and in the transition rates for the binary components that are accounted for by the Fisher Information Matrix of Section 2.1.2. The outmost loop within the following procedure (sketched also in Figure 14) consists in randomly sampling the values of the physical RPS model parameters from their distributions and sampling the RTD failure time (step 4):

1) Set initial time t_0=0, mission time T=6yr and time step dt=0.01yr;

2) Randomly sample the transition rates of the binary-states components (i.e., BPLs, LCLs, and RTB) from the Gaussian distributions $N\left(\lambda_k, \hat{\sigma}_k\right)$ of Table 4, where, k = B, L, R;

3) Sample the failure times of the binary-states components, from the exponential distributions with the sampled transition rates;

4) Randomly sample the multi-state RTD failure time by:

    4a) Fit the randomly sampled realizations of the RTD response time τ at each discrete point to a trial surface $\tau(t,\delta)$;

4b) Simulate the RTD degradation process evolution from $t=t_0$ to $t=T$;

4c) At each time t, sample the air gap size increment $d\delta_t$ from a normal distribution $N(0,0.025t)$, resulting in $\delta = \delta_0 + d\delta_t$.

4d) Calculate the response time τ on the fitted trial surface $\tau(t,\delta)$.

4e) Record the time t at which τ exceeds the threshold γ_Y=8s, with air gap size δ.

5) Integrate the RTD-MSPM into the RPS-MSPM;

5a) Sort all the components sampled failure times;

5b) Check whether the minimum of the sorted times exceeds T:

- If yes, increase the unreliability counter at time T;
- If not, check whether at that time the RPS-MSPM reaches any absorbing state and, if yes, increase the unreliability counter, or the reliability counter, otherwise.

6) Run Nb=1000 times steps 1) to 5) to build the empirical $P(t|\delta;\lambda_B;\lambda_L;\lambda_R)$, based on the statistics of the system unreliability estimates collected at each time t;

7) Estimate the 5th and 95th percentiles of the unreliability by repeating steps 1)-6) for Nc=1000 times and collecting the related statistics;

8) Obtain the pointwise double-sided 90% confidence intervals of the system unreliability calculated by the RPS-MSPM.

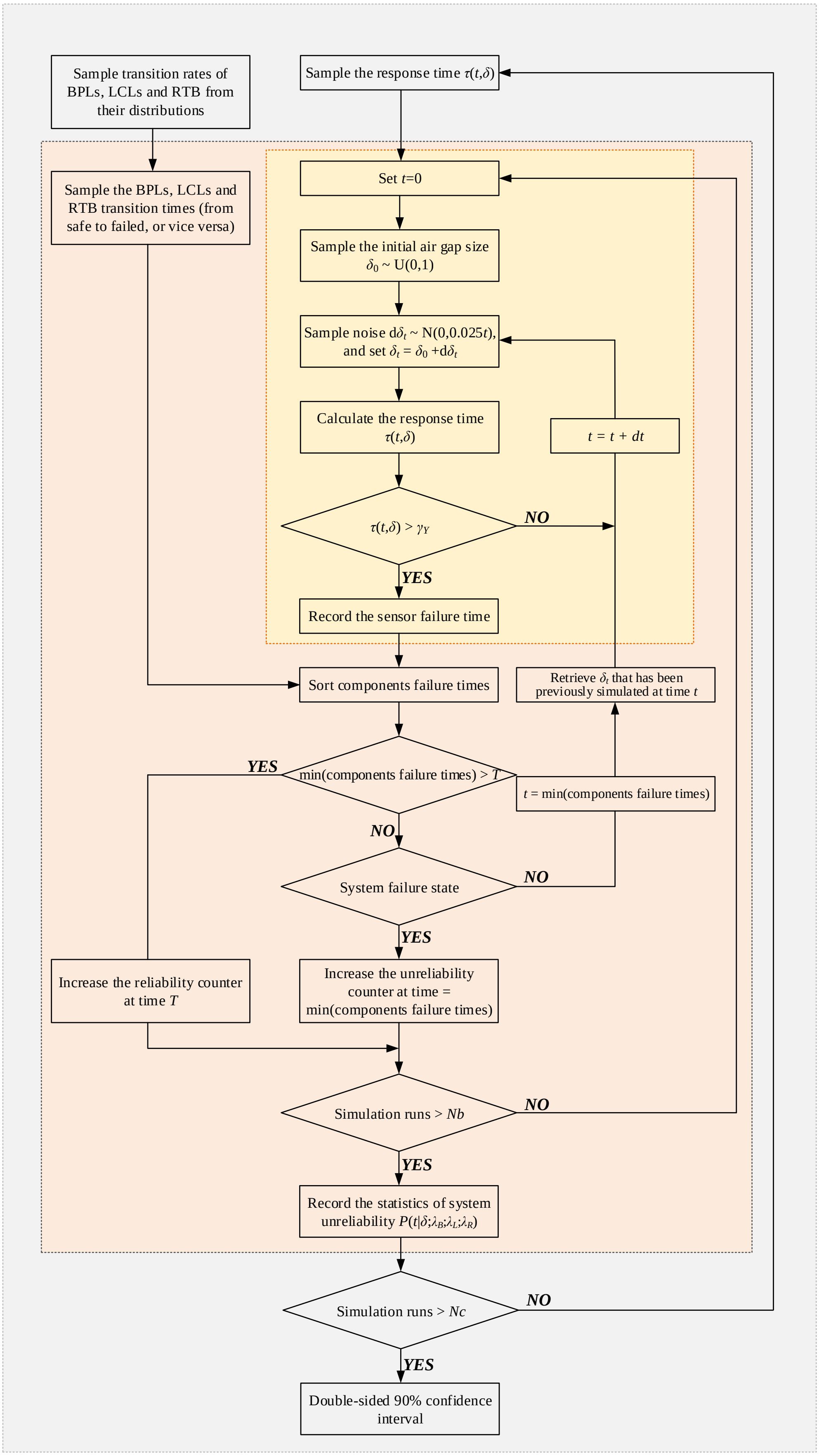

Sample transition rates of BPLs, LCLs and RTB from their distributions
Sample the response time $\tau(t,\delta)$
Sample the BPLs, LCLs and RTB transition times (from safe to failed, or vice versa)
Set t=0
Sample the initial air gap size δ_0 ~ U(0,1)
Sample noise dδ_t ~ N(0,0.025t), and set $\delta_t = \delta_0$ +dδ_t
Calculate the response time $\tau(t,\delta)$
$t = t + dt$
$\tau(t,\delta) > \gamma_Y$
NO
YES
Record the sensor failure time
Sort components failure times
Retrieve δ_t that has been previously simulated at time t
YES
min(components failure times) > T
t = min(components failure times)
NO
System failure state
NO
YES
Increase the reliability counter at time T
Increase the unreliability counter at time = min(components failure times)
Simulation runs > Nb
NO
YES
Record the statistics of system unreliability $P(t|\delta;\lambda_B;\lambda_L;\lambda_R)$
Simulation runs > Nc
NO
YES
Double-sided 90% confidence interval

Figure 14 Flowchart of the three-loop MC simulation

Figure 15 shows the estimated P(t,δ) with the 90% confidence interval for the RPS-MSPM, obtained by the three-loop MC simulation. The confidence interval is large especially in [0.5, 1.5] yr, probably because the fitted trial surfaces at the basis of the uncertainty propagation considerably vary from each other due to the large variances of the data of Tables 7 and 8 utilized to build them. Despite that, as we shall see in what follows, the robustness of the assessment is much improved with respect to the RPS-MCM results.

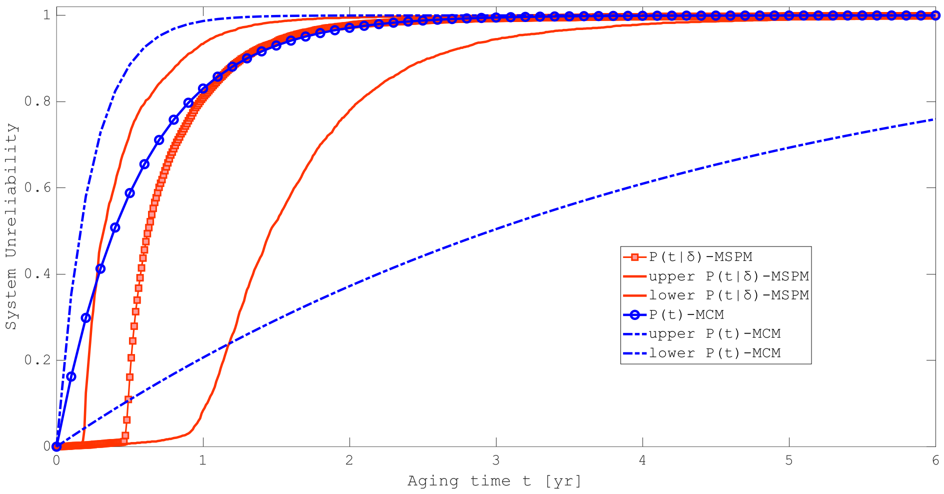


Figure 15 Estimated RPS-MSPM unreliability with 90% confidence interval

2.5. Comparison of the RPS-MSPM estimates with the RPS-MCM estimates

Figure 15 also shows the results of the RPS reliability assessment by the RPS-MCM that does not take into account the RTD degradation-to-failure process. In general terms, it can be concluded that the RPS-MSPM results provide a narrower confidence interval than the RPS-MCM, thanks to the integration of physics knowledge related to operational and environmental parameters. The confidence interval provided by the MSPM is larger than that of the MCM at the early stage of the RPS life (t<1yr): the main reason is that the fitting surfaces may considerably vary from trial to trial due to the large variance of the response times at the considered discrete points, which greatly affect the onset time of the RTD drift failure mode.

For a quantification analysis, two indexes (i.e., the relative uncertainty interval width ζ_t and the relative age interval width ζ_P) are proposed in what follows to compare the accuracy of the MCM with that of the MSPM.

(1) The relative uncertainty interval width

At each time t, the ratio ζ_t between the mean value of the system unreliability and the width of the unreliability interval (i.e., the difference between the upper and lower bounds) is calculated.

The larger ζ_t, the narrower is the confidence interval, and the more accurate the system reliability modeling approach. Figure 16 shows that $\zeta_t(t|\delta)$ of the MSPM is much larger than $\zeta_t(t)$ of the MCM: as t increases, the estimated system unreliability obviously increases but, since MSPM includes more (physics) knowledge on the system behavior than MCM, the confidence interval reduces more than that of the MCM. The zoom of Figure 16 shows the evolution of ζ_t from t=0 to t=2yr: to further investigate the dispersion of the unreliability estimates within the bounds, we calculate, at each time, their empirical pdf and the respective cdf.

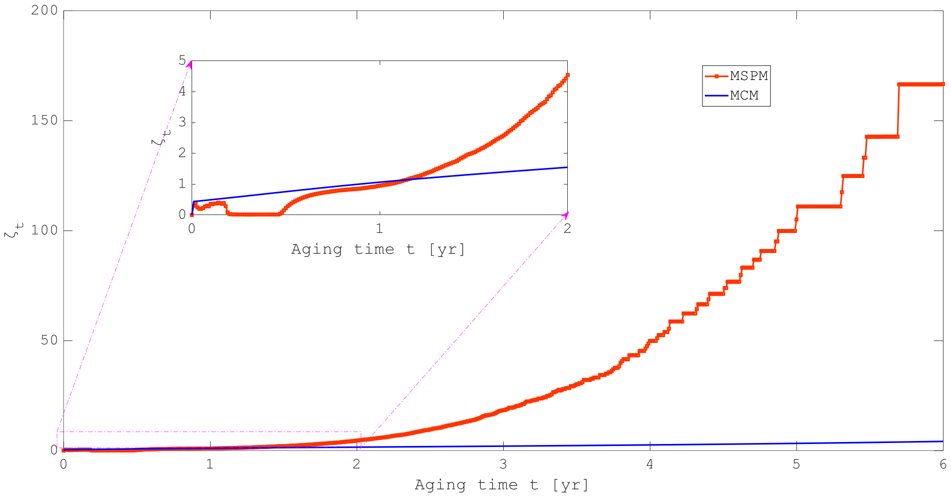


Figure 16 Relative unreliability interval width

Based on the real estimates collected with the *Na* MC simulations for the RPS-MCM reliability assessment of Section 2.1.3 and the *Nc* three-loop MC simulation for the RPS-MSPM, Figures 17 shows the pdf curve of the system unreliability at t=1*yr*. The pdf of the MCM skews towards large unreliability values, compared to the pdf of the MSPM, demonstrating again the more probable overestimation of the system unreliability, if the decision maker were to resort to RPS-MCM.

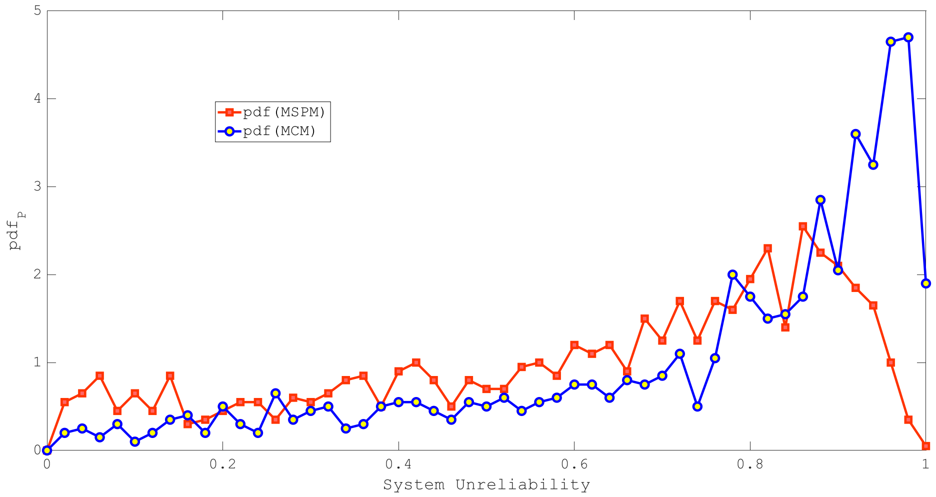


Figure 17 pdf of the MCM estimates vs. pdf of the MSPM estimates at t=1yr

(2) The relative age interval width

With respect to each system reliability value *P*, the ratio ζ_P between the mean value of the system failure time and the width of the age interval (i.e., the difference between the upper and lower bounds), is calculated.

The larger ζ_P, the narrower the confidence interval, and the more accurate the system reliability estimate. Figure 18 shows $\zeta_P(P|\delta)$ of the MSPM and $\zeta_P(P)$ of the MCM. The latter is always larger than the former, whatever the value of *P*, that means that MSPM better models the RTD degradation and, therefore, provides more accurate failure time predictions than the MCM. For clarity sake, $\zeta_P(P)$ of the MCM is truncated at *P*=0.8 because the maximum unreliability of the lower bound of the MCM is 0.8 within the mission time.

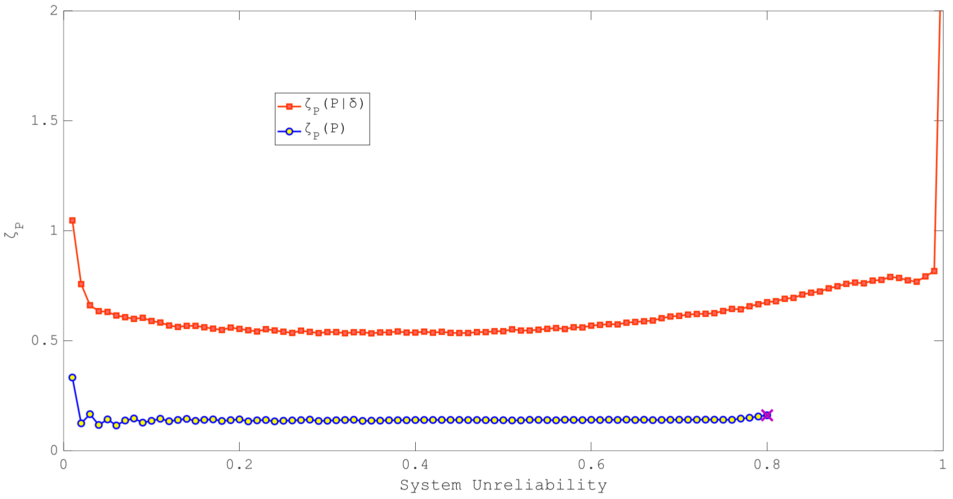


Figure 18 Relative age interval width

To further investigate the dispersion of the age interval estimates, we calculate, at each unreliability value *P*, the empirical pdf and respective cdf. Resorting to the real estimates collected with the *Na* MC simulation for the RPS-MCM reliability assessment of Section 2.1.3 and the *Nc* three-loop MC simulation for the RPS-MSPM of Section 2.4, Figures 19 is built with the pdf curve of the system failure times at *P*=0.1, respectively. The pdf of the MCM skews towards the earlier values, compared with the pdf of the MSPM, revealing the more possible early-estimation of the failure times, if the decision maker resorts to a MCM.

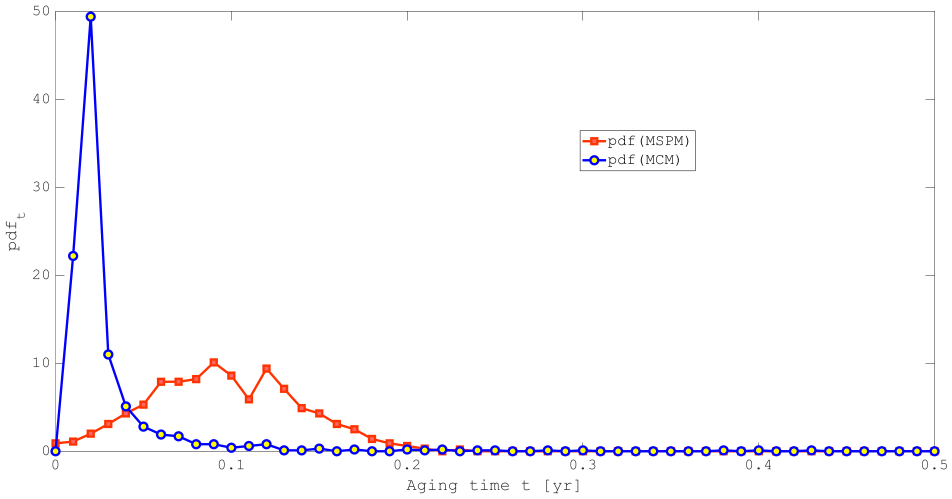


Figure 19 pdf of the MCM estimates vs. pdf of the MSPM estimates at P=0.1

2.6. Conclusions

In this Chapter, a system-level MSPM model has been proposed for the safety assessment of CPSs, where the MSPM offers the possibility of embedding the physical degradation process into the safety assessment of the systems. The methods proposed leverage the demanding efforts needed for modeling the physical relationships and the high computational burden of manipulating the large amount of data, especially when treating uncertainties.

In practice, a SA has been performed in Section 2.2, to identify the components of a system that most deserve accurate modeling of aging- and environmental-dependent transition rates, for accurate system reliability assessment and for trading model accuracy and computational demand for practical reliability assessment based on MSPM. The SA has been performed based on moment-independent sensitivity measures, such as Hellinger distance and Kullback-Leibler divergence. Application to a RPS of NPPs, the SA has led to focusing on the reliability assessment of a RTD, which is an important digital I&C component used to guarantee the safe operation of NPPs (Paper 1[C]). On this result, a MSPM has been built in Section 2.3 to describe this component degradation towards failure and MC simulation has been used to estimate the probability of sojourn in any of the degradation states (Paper 1[J]). The resulting model has, then, been integrated into a system-level MSPM of the RPS, to estimate the system failure probability accounting for both aging- and environmental-dependent transition rates of the RTD (thus, embedding more knowledge into the modeling of the most important contributors to the RPS unreliability, compared with traditional dynamic methods, e.g., MCM) (Papers 1[J]). In Section 2.4, a three-loop MC simulation scheme has been developed to operationalize the MSPM approach for large scale systems, and to quantify and control the uncertainty affecting the system reliability model (Papers 2[J]).

Results from a nuclear CPS case study show the novel methods can give due account to uncertainties affecting aging, degradation and stochastic failures of CPS components in hazard analysis.

3. MODELING AND SIMULATION FOR THREAT ANALYSIS

Contents of the Chapter have been adapted from:

3[J] Wang, W., Cammi, A., Di Maio, F., Lorenzi, S. and Zio, E., 2018. A Monte Carlo-based exploration framework for identifying components vulnerable to cyber threats in nuclear power plants. Reliability Engineering & System Safety, 175, pp.24-37.

CPSs functionality can be compromised also by security breaches (such as cyber attacks). Multiple failure modes (such as bias, drift and freezing) can occur, both due to random failures or induced by malicious external attacks. In this Chapter, we illustrate an exploration approach that, based on safety margins estimation, allows identifying the most vulnerable components to malicious external attacks. For demonstration, we apply the approach to the ALFRED. Its object-oriented model is embedded within a MC-driven engine that injects different types of cyber attacks at random times and magnitudes. Safety margins are, then, calculated and used for identifying the most vulnerable CPS components. This allows selecting protections to make ALFRED resilient towards maliciously induced failures.

3.1. Case study: the ALFRED

The ALFRED reactor with its full power mode control scheme and the MC engine of cyber breaches injection are described in Sections 3.1.1 and 3.1.2, respectively.

3.1.1. The ALFRED full power mode control scheme

At full power nominal conditions, the dynamics processing of the ALFRED primary and secondary cooling systems is controlled by a multi-loop PI control scheme, as shown in Figure 20. The PI-based feedback control configuration employs four SISO control loops independent of each other. The parameters specification of ALFRED at full power nominal conditions are reported in Table 9.

Table 9 ALFRED parameters values, at full power nominal conditions

Parameter	Parameter Description	Value	Unit
P_{Th}	Thermal power	$300\cdot10^6$	W
h_{CR}	Height of control rods	12.3	cm
$T_{L,hot}$	Coolant core outlet temperature	480	°C
$T_{L,cold}$	Coolant Steam Generator (SG) outlet temperature	400	°C
Γ	Coolant mass flow rate	25984	kg·s^{-1}
T_{feed}	Feedwater SG inlet temperature	335	°C
T_{steam}	Steam SG outlet temperature	450	°C
p_{SG}	SG pressure	$180\cdot10^5$	Pa
G_{water}	Feedwater mass flow rate	192	kg·s^{-1}
G_{att}	Attemperator mass flow rate	0.5	kg·s^{-1}
kv	Turbine admission valve coefficient	1	-
P_{Mech}	Mechanical power	$146\cdot10^6$	W

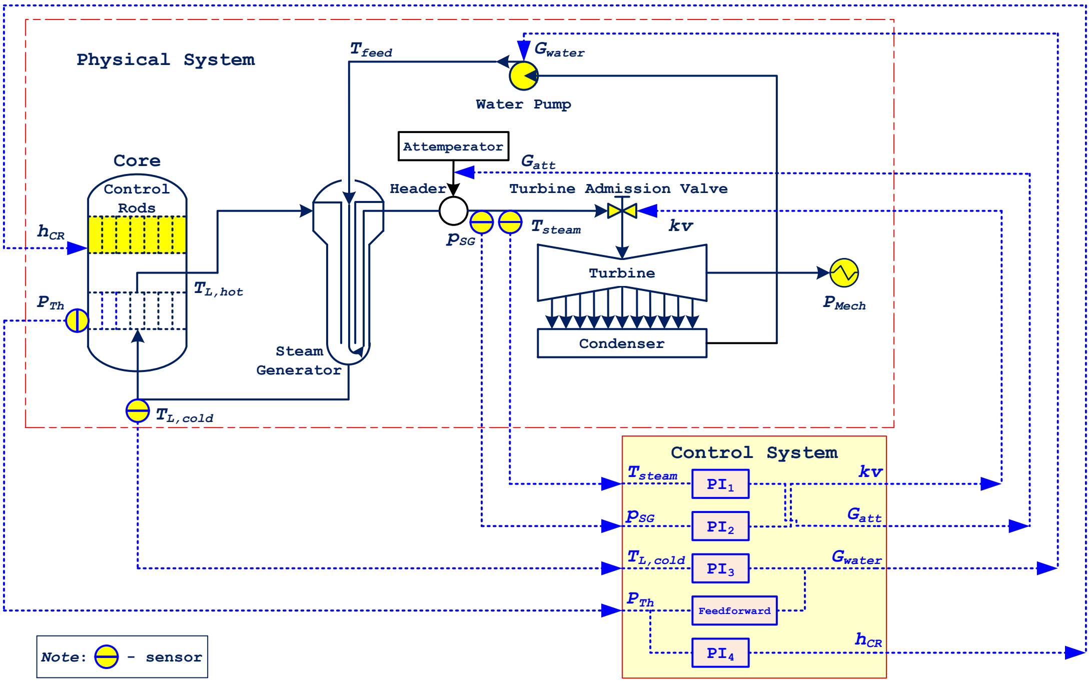


Figure 20 ALFRED reactor control scheme

The control aims at keeping the controlled variables of the control loops approximately at the steady state values, for outputting a steady mechanical power. The values represent the optimal working conditions of the system at full power nominal conditions. The regulation of the controlled variables is of particular concern, to bring benefits to the structural materials and ensure safe NPP operation conditions. Safety thresholds for each variable, listed in Table 10, are set such that consequences of transients and accidents are limited: for example, the $T_{L,cold}$ must be kept above 350°C to avoid the embrittlement of the structural materials in aggressive environments enhanced by the fast neutron irradiation.

In Figure 21, profiles of the controlled variables, with a mission time t_M equal to 3000s, are shown. Under the control scheme of Figure 20, the values of the variables are kept approximately at their nominal values, at full power nominal conditions, despite the measuring errors (white noise).

Table 10 List of reference and threshold values for safety variables

Variable, *y*	Reference value, R_y, at full power nominal conditions	Safety thresholds	
		Lower, L_y	Upper, U_y
T_{steam} (°C)	450	-	550
p_{SG} (Pa)	$180 \cdot 10^5$	$170 \cdot 10^5$	$190 \cdot 10^5$
$T_{L,cold}$ (°C)	400	350	-
P_{Th} (W)	$300 \cdot 10^6$	$270 \cdot 10^6$	$330 \cdot 10^6$

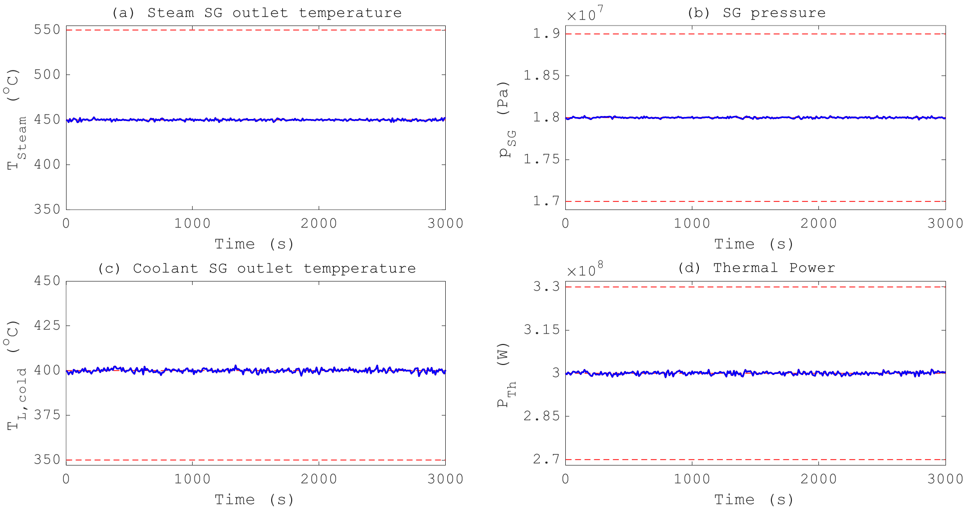


Figure 21 Profiles of the controlled variables of the ALFRED model at full power nominal conditions: (a) Steam SG outlet temperature; (b) SG pressure; (c) Coolant SG outlet temperature; and (d) Thermal power

3.1.2. The Monte Carlo engine of cyber breaches injection

To test the effects of cyber attacks on system integrity, a MC engine is integrated with the ALFRED model for injecting cyber breaches at random times and magnitudes. It shall be noted that, the random time t_R of the attack occurrence only plays an illustrative role in modeling the random occurrence of a cyber attack in reality. The cyber attacks here considered are sketched in Figure 22 and hereafter described.

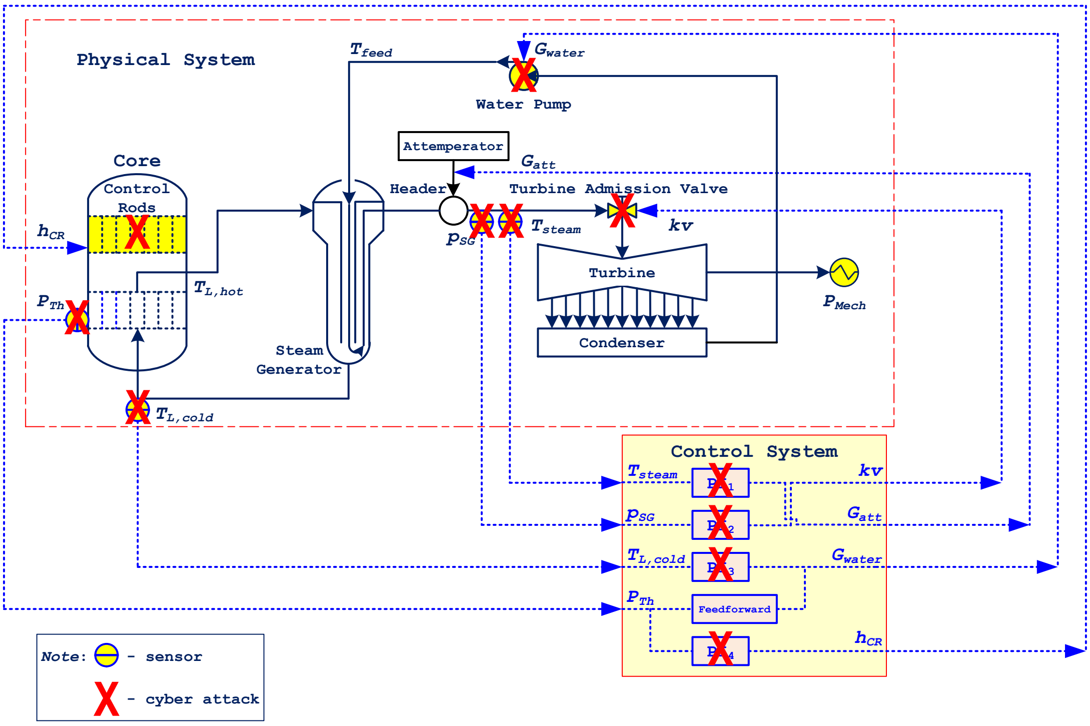


Figure 22 Sketch of cyber attacks injected into the ALFRED system

(1) Sensors

Four types of cyber attacks occurring at random time t_R are considered for each sensor, preventing the controllers from receiving legitimate measurements (equivalent to typical DoS attacks [83, 201, 202]), mimicking stochastic failures [203]: (*a*) bias, (*b*) drift, (*c*) wider noise and (*d*) freezing. The occurrence of any of these failure modes results in altered sensor measurements $y_{sensor}(t)$, as in Eq. (3-1):

$$y_{sensor}(t) = \begin{cases} y(t)+\delta(t), & \delta(t) = N(0,\sigma), \sigma > 0, & t \geq 0, & \text{normal} \\ y(t)+\delta(t)+b, & \dot{b}(t) \equiv 0, b(t_R) \neq 0, & t \geq t_R, & \text{bias} \\ y(t)+\delta(t)+c(t), & c(t) = c \cdot (t - t_R), & t \geq t_R, & \text{drift} \\ y(t)+\delta'(t), & \delta'(t) = N(0,\alpha\sigma), \alpha > 1, & t \geq t_R, & \text{wider noise} \\ y_{sensor}(t_R), & & t \geq t_R, & \text{freezing} \end{cases} \qquad (3\text{-}1)$$

where $y(t)$ is the real value of the controlled variable y at time t, $\delta(t)$ is the nominal measuring error, distributed according to a normal distribution $N(0,\sigma)$, b is a constant bias factor, c is a constant drift factor, $\delta'(t)$ is a wider measuring error, distributed according to a normal distribution $N(0,\alpha\sigma)$ with a larger variance than $\delta(t)$ (α>1).

Practically, the MC sampling procedure used to inject a random cyber attack to sensors at time t_R consists in sampling the uncertain parameters b, c, $\delta'(t)$ from the distributions listed in Table 11 and, then, running the ALFRED simulator for collecting the controlled variables evolution throughout the mission time t_M. Notice that Gaussian noises are typical of sensor data acquisition, leading to sensor nominal errors (column 2) and wider errors (column 5) under nominal condition and wider noise failure mode, respectively. Bias and drift (columns 3 and 4, respectively) are, instead, a-priori set from uniform distributions, to mimic sensor stochastic failures due to cyber attacks.

Table 11 Parameters of sensors

Sensor	Nominal error $\delta(t)$	Failure factors		
		Bias b	Drift c	Wider noise $\delta'(t)$
T_{steam} (°C)	$N(0,1)$	$U(-200,200)$	$U(-1,1)$	$N(0,10)$
p_{SG} (Pa)	$N(0,0.1) \cdot 10^5$	$U(-100,30) \cdot 10^5$	$U(-0.2,0.2) \cdot 10^5$	$N(0,2) \cdot 10^5$
$T_{L,cold}$ (°C)	$N(0,1)$	$U(-30,30)$	$U(-1,1)$	$N(0,5)$
P_{Th} (W)	$N(0,0.5) \cdot 10^6$	$U(-300,30) \cdot 10^6$	$U(-0.5,0.5) \cdot 10^6$	$N(0,0.7) \cdot 10^6$

(2) Actuators

Three actuators of the digital I&C system of ALFRED are considered susceptible of a malicious attack, namely: control rods that regulate the rod heights h_{CR}, water pump that regulates the feedwater mass flow rate G_{water} and turbine admission valve kv that regulates the steam inlet mass flow rate. At nominal conditions, the actuators execute the command signals of the control system to respond to the sensors measurements and accommodate disturbances, transients or accidents. On the other hand, under attack, the actuators might fail stuck to a random magnitude of actuation $A(t)$, here sampled from a uniform distribution (see Table 12): in this situation, the actuators would no longer receive proper control commands and the I&C system would not be capable of accommodating disturbances, transients or accidents.

Table 12 Parameters of actuators

Actuator	Regulated control variable	Reference regulation	Failure distribution
Control rods	h_{CR} (cm)	12.3	U(0,64)
Water pump	G_{water} (kg·s^{-1})	192	U(0,300)
Turbine admission valve coefficient	kv (-)	1	U(1,1.5)

(3) PI controllers

At nominal conditions, PI gains (i.e., K_p and K_i) and controlled variables set points $y_{set,ref}$ are fixed by the control designers, to keep the physical process variables close to their nominal values. Under the cyber attack, equivalent to a deception attack maliciously injecting a false message to the controller [50, 201], PI gains and set points are randomly sampled from uniform distributions, covering all possible values (see Table 13). In terms of uniform distributions for sampling random values of PI gains (columns 6 and 7), their expectations are larger than the reference values, for increasing the possibility that the cyber attack impacts the system integrity [204].

Table 13 Parameters of PIs

PI	Controlled variable, y	Reference value			PI parameter upon attack		
		$K_{p,ref}$	$K_{i,ref}$	Set point, $y_{set,ref}$	K_p	K_i	Set point, y_{set}
PI_1	T_{steam}	$1 \cdot 10^{-1}$	$5 \cdot 10^{-2}$	450 (°C)	$U(1 \cdot 10^{-2},1)$	$U(5 \cdot 10^{-4},5)$	U(430,470) (°C)
PI_2	p_{SG}	$3 \cdot 10^{-7}$	$1 \cdot 10^{-8}$	$180 \cdot 10^5$ (Pa)	$U(3 \cdot 10^{-8},3 \cdot 10^{-4})$	$U(3 \cdot 10^{-10},3 \cdot 10^{-5})$	$U(170,190) \cdot 10^5$ (Pa)
PI_3	$T_{L,cold}$	$6 \cdot 10^{-1}$	$1 \cdot 10^{-2}$	400 (°C)	$U(6 \cdot 10^{-2},6)$	$U(1 \cdot 10^{-4},1)$	U(380,420) (°C)
PI_4	P_{Th}	$2 \cdot 10^{-11}$	$4 \cdot 10^{-11}$	$300 \cdot 10^6$ (W)	$U(2 \cdot 10^{-12},2 \cdot 10^{-7})$	$U(4 \cdot 10^{-13},4 \cdot 10^{-6})$	$U(285,315) \cdot 10^6$ (W)

It is worth mentioning that the components of the digital I&C system considered, their failure modes and cyber attack types are not intended to provide a comprehensive description of the system accidental behavior, but are only taken as exemplary for generating the dynamic accident scenarios to be processed for safety margins estimation, within the framework here proposed for the identification of the components most vulnerable to cyber threats. Moreover, we observe that an attacker is interested also in injecting "soft" failures that slowly drive the system into failure, rather than, only "hard" failures because the former is more difficult to detect and recover.

3.2. Cyber threats prioritization

In this Section, a safety margin estimation approach of literature [155-157] is utilized for cyber threat prioritization. It is here originally undertaken for processing cyber attack scenarios, to estimate the extent of the consequences of cyber threats on the CPS components.

3.2.1. The safety margins estimation approach for cyber threats prioritization

(1) One-sided safety margin

Considering a set of accidental scenarios *a* simulated over a mission time t_M, the safety margin a safety parameter *y*, with respect to a predefined upper threshold U_y (see Figure 23) is defined as the ratio between the computed value reached by the value of a specific γ_1 percentile of the distribution of the measured maximum values, $y^{\gamma_1}_{max,a}$ and the design value y_{ref}, where $\hat{y}^{\gamma_1,\beta_1}_{max,a}$ (i.e., the estimate of $y^{\gamma_1}_{max,a}$) is given with confidence β_1 [205], viz:

$$\begin{cases} \gamma_1 = \Pr\left(y_{\max,a} < y^{\gamma_1}_{\max,a}\right) \\ \beta_1 = \Pr\left(y^{\gamma_1}_{\max,a} < \hat{y}^{\gamma_1,\beta_1}_{\max,a}\right) \end{cases} \tag{3-2}$$

and,

$$M^{\gamma_1,\beta_1}_{U,y_a} = \begin{cases} \dfrac{U_y - \hat{y}^{\gamma_1,\beta_1}_{\max,a}}{U_y - y_{ref}} & y_{ref} < \hat{y}^{\gamma_1,\beta_1}_{\max,a} < U_y \\ 0 & U_y \le \hat{y}^{\gamma_1,\beta_1}_{\max,a} \\ 1 & \hat{y}^{\gamma_1,\beta_1}_{\max,a} \le y_{ref} \end{cases} \tag{3-3}$$

The value $\hat{y}^{\gamma_1,\beta_1}_{\max,a}$ is estimated by the Bracketing OS approach, which allows controlling the computational cost of the simulation codes and guarantees that the first element (out of *N*) in the descending sorted sample $y^1_{\max,a}$ has a certain probability β_1 of exceeding the unknown true γ_1 percentile. The number *N* can be calculated by Eq. (3-4), when γ_1 and β_1 are predefined.

$$\beta_1 = 1 - \gamma_1^N \tag{3-4}$$

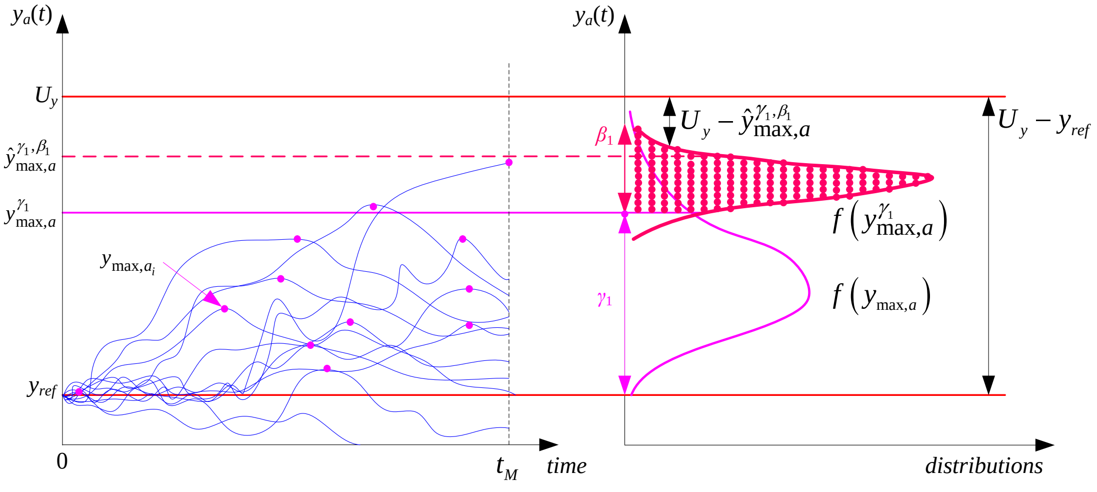

$y_a(t)$
U_y
$\hat{y}^{\gamma_1,\beta_1}_{\max,a}$
$y^{\gamma_1}_{\max,a}$
$y_{\max,a_i}$
y_{ref}
0
t_M
time
$y_a(t)$
$U_y - \hat{y}^{\gamma_1,\beta_1}_{\max,a}$
β_1
$U_y - y_{ref}$
$f\left(y^{\gamma_1}_{\max,a}\right)$
γ_1
$f\left(y_{\max,a}\right)$
distributions

Figure 23 $y_{\max,a_i}$ obtained from N samples of the accidental scenario a used to estimate $\hat{y}^{\gamma_1,\beta_1}_{\max,a}$, and, thus, to estimate $M^{\gamma_1,\beta_1}_{U,y_a}$

Similarly, the safety margin with respect to a lower threshold L_y becomes:

$$M_{L,y_a}^{\gamma_2,\beta_2} = \begin{cases} \dfrac{\hat{y}_{\min,a}^{\gamma_2,\beta_2} - L_y}{y_{ref} - L_y} & L_y < \hat{y}_{\min,a}^{\gamma_2,\beta_2} < y_{ref} \\ 0 & \hat{y}_{\min,a}^{\gamma_2,\beta_2} \le L_y \\ 1 & y_{ref} \le \hat{y}_{\min,a}^{\gamma_2,\beta_2} \end{cases} \tag{3-5}$$

where, $\hat{y}_{\min,a}^{\gamma_2,\beta_2}$ is the point estimate value of the γ_2 percentile of the distribution of the measured values $y_{\min,a}$, with a confidence β_2, and, γ_2 and β_2 are:

$$\begin{cases} \gamma_2 = \Pr\left(y_{\min,a} < y_{\min,a}^{\gamma_2}\right) \\ \beta_2 = \Pr\left(y_{\min,a}^{\gamma_2} > \hat{y}_{\min,a}^{\gamma_2,\beta_2}\right) \end{cases} \tag{3-6}$$

The number N can be calculated by Eq. (3-7), when γ_2 and β_2 are predefined.

$$\beta_2 = 1 - \left(1 - \gamma_2\right)^N \tag{3-7}$$

(2) Two-sided safety margin

The safety margin M_{T,y_a} of a safety parameter y with respect to the double-sided (both upper U_y and lower L_y) thresholds (see Figure 24) is defined as the minimum value between $M_{U,y_a}^{\gamma_1,\beta}$ and $M_{L,y_a}^{\gamma_2,\beta}$ of Eqs. (3-3) and (3-5):

$$M_{T,y_a}^{\gamma_1,\gamma_2,\beta} = \min\left(M_{U,y_a}^{\gamma_1,\beta}, M_{L,y_a}^{\gamma_2,\beta}\right) \tag{3-8}$$

where, the number of the scenario samples N to be sorted can be calculated, when γ_1, γ_2 and β are predefined (Nutt and Wallis, 2004), according to Eq. (3-9):

$$\beta = 1 - \gamma_1^N - \gamma_2^N + \left(\gamma_1 + \gamma_2 - 1\right)^N \tag{3-9}$$

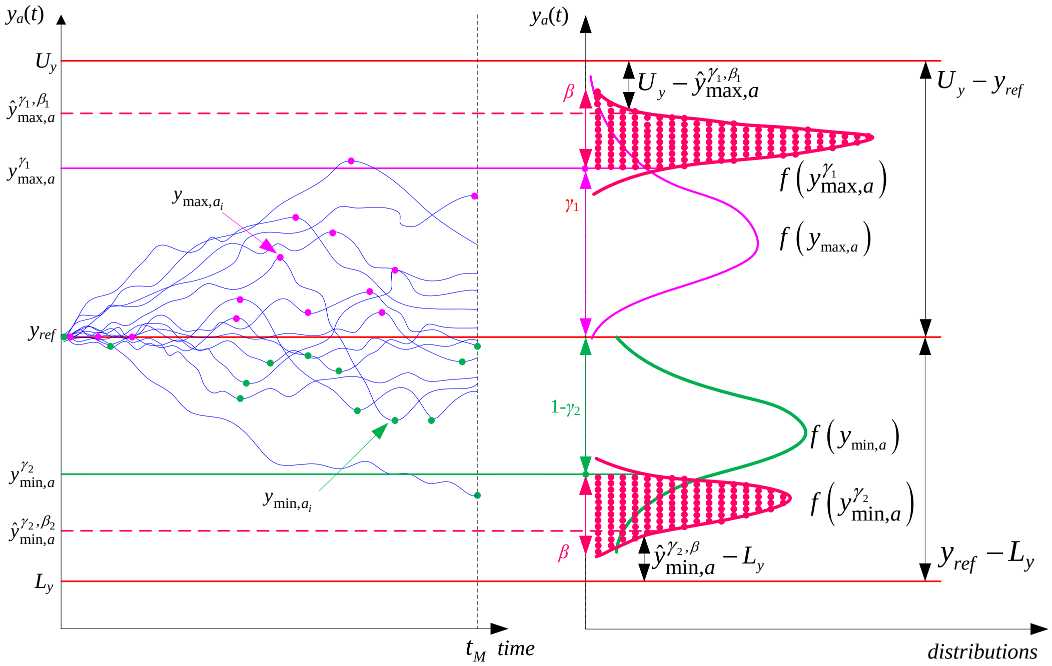

$y_a(t)$
U_y
$\hat{y}^{\gamma_1,\beta_1}_{\max,a}$
$y^{\gamma_1}_{\max,a}$
$y_{\max,a_i}$
y_{ref}
$y_{\min,a_i}$
$y^{\gamma_2}_{\min,a}$
$\hat{y}^{\gamma_2,\beta_2}_{\min,a}$
L_y
t_M time
$y_a(t)$
β
γ_1
1-γ_2
β
$U_y - \hat{y}^{\gamma_1,\beta_1}_{\max,a}$
$U_y - y_{ref}$
$f\left(y^{\gamma_1}_{\max,a}\right)$
$f\left(y_{\max,a}\right)$
$f\left(y_{\min,a}\right)$
$f\left(y^{\gamma_2}_{\min,a}\right)$
$\hat{y}^{\gamma_2,\beta}_{\min,a} - L_y$
$y_{ref} - L_y$
distributions

Figure 24 N pairs of maximum and minimum values of the accidental scenario are used to estimate $\hat{y}^{\gamma_1,\beta}_{\max,a}$ and $\hat{y}^{\gamma_2,\beta}_{\min,a}$, and, thus, to estimate $M^{\gamma_1,\gamma_2,\beta}_{T,y_a}$

The responses of ALFRED to cyber attacks to sensors, actuators and PI regulators are investigated by simulation. From the simuations outcomes, safety margins of the four controlled variables y (i.e., steam outlet temperature of Steam Generator (SG) T_{steam}, SG pressure p_{SG}, coolant SG outlet temperature $T_{L,cold}$, and thermal power P_{Th}) are estimated, to quantify the effects of the cyber attacks on the system functionalities. A total of N_T=29 runs of the ALFRED model are simulated, to satisfy the requirements of the percentiles estimations of the safety parameters by the Bracketing OS, with respect to both one-sided (N=22, given *i*) γ_1= 90th, β_1=90th, or *ii*) γ_2=10th, β_2=90th) and two-sided (N=29, given γ_1,= 90th, γ_2=10th, β=90th) thresholds. Accordingly, N=22 samples are randomly taken to estimate the safety margins of T_{steam} (with respect to its upper threshold) and $T_{L,cold}$ (with respect to its lower threshold) and N=29 samples are used to estimate the safety margins of p_{SG} and P_{Th} (with respect to their two-sided thresholds).

Effects of cyber attacks on the CPS components and on the system integrity are qualitatively ranked according to a three-level risk metric (see Table 14). Table 15 shows the quantified design safety margins when the code is run 29 times under nominal conditions (proving that the system works with ample safety margins).

Table 14 A three-level risk metric for ranking the effects of cyber attacks on the CPS

Effect	Safety magin $M_{\#3}^{\#1(\text{or }\#2)}$
Negligible	[0.8, 1.0]
Medium	[0.2, 0.8)
Severe	[0.0, 0.2)

Note: 1) #1 refers to "γ_2,β_2" for $T_{L,cold}$, and to "γ_2,β" for p_{SG} and P_{Th};
2) #2 refers to "γ_1,β_1" for T_{steam}, and to "γ_1,β" for p_{SG} and P_{Th};
3) #3 refers to "U,y_a" for T_{steam}, to L,y_a for $T_{L,cold}$, and to "T,y_a" for p_{SG} and P_{Th};
4) γ_1= 90th, γ_2=10th, β_1=90th, β_2=90th, β=90th.

Table 15 Safety margins estimation of the safety parameters under normal conditions

Variable	T_{steam} (°C)	p_{SG} (Pa)	$T_{L,cold}$ (°C)	P_{Th} (W)
$\hat{y}_{\min,y_a}^{\#1}$	-	$1.7967 \cdot 10^7$	396.1839	$2.9819 \cdot 10^8$
$\hat{y}_{\max,y_a}^{\#2}$	455.3330	$1.8029 \cdot 10^7$	-	$3.0181 \cdot 10^8$
$M_{\#3}^{\#1(\text{or }\#2)}$	0.9667	0.9672	0.9237	0.9396

3.3. Results

(1) Sensors

Table 16 presents the results of the safety margins estimation of the four types of failure modes of the four sensors measuring the values of the controlled variables, i.e., T_{steam}, p_{SG}, $T_{L,cold}$, and P_{Th}.

Table 16 Safety margins estimation of the safety parameters of the cyber attacks to sensors

Scenario *a*		$M_{U,T_{steam,a}}^{\gamma_1,\beta_1}$	$M_{T,p_{SG,a}}^{\gamma_1,\gamma_2,\beta}$	$M_{L,T_{L,cold,a}}^{\gamma_2,\beta_2}$	$M_{T,P_{Th,a}}^{\gamma_1,\gamma_2,\beta}$
T_{steam} sensor	bias	0.9562	0.9626	0.9350	0.9349
	drift	0.9604	0.9654	0.9188	0.9322
	wider noise	0.9579	0.9478	0.9195	0.9330
	freezing	0.9604	0.9654	0.9203	0.9374
p_{SG} sensor	bias	0.8185	0	0.7776	0.8136
	drift	0.8701	0	0.9042	0.9335
	wider noise	0.9349	0.4098	0.9257	0.9220
	freezing	0.8988	0	0.9031	0.8600
$T_{L,cold}$ sensor	bias	0.5875	0.5787	0.5436	0.3838
	drift	0	0	0.5469	0
	wider noise	0.9138	0.9002	0.9085	0.9073
	freezing	0.2187	0.9722	0.6261	0.4707
P_{Th} sensor	bias	0.9641	0	0.2342	0
	drift	0.9539	0.9662	0.8811	0
	wider noise	0.9601	0.9649	0.9212	0.9326
	freezing	0.9657	0.9645	0.9261	0.7672

The results show that cyber attacks leading to T_{steam} sensor failures do not affect the system functioning because all safety parameters are negligibly affected. System integrity can be affected by cyber attacks to the p_{SG}, $T_{L,cold}$ and P_{Th} sensors, directly resulting in large variations of the respective variables (attacks to P_{Th} with a minor impact on the other controlled variables, whereas, cyber attacks to $T_{L,cold}$ sensor, e.g., bias, drift, or freezing, may impact the whole physical system).

As example, Figure 25 shows the evolution of the safety parameters when the $T_{L,cold}$ sensor is affected by the freezing failure mode. In all cases, the lead temperature at the SG outlet, $T_{L,cold}(t)$ deviates from its set point equal to 400°C (Figure 25(a)), due to the PI_3 response to the frozen value $T_{L,cold,sensor}(t)$. Then, the steam SG outlet temperature T_{steam} changes accordingly to the change of the lead temperature (Figure 25(b)), causing the change of Thermal power P_{Th} (Figure 25(d)). SG pressure change (Figure 25(c)) is negligible thanks to the effective regulation of the steam mass flow rate by the turbine admission valve. These alterations are well caught by the safety margin analysis. In particular, the safety margin of T_{steam}, $T_{L,cold}$, and P_{Th} in case of $T_{L,cold}$ sensor freezing (Table 17) result to be equal to 0.2187, 0.6260, and 0.4707, respectively. This corresponds to a "medium" effect, according to the predefined risk metric of Table 14. On the other hand, SG *Pressure* is kept approximately at the nominal level with little disturbances, and, thus, "negligibly" affected by the cyber attacks.

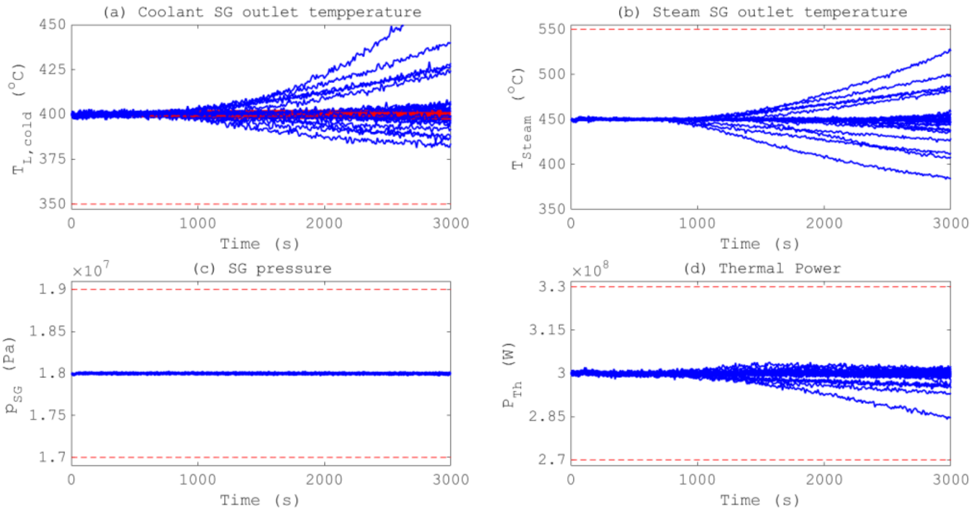


Figure 25 Profiles of the safety parameters for N_T=29 runs, when $T_{L,cold}$ sensor is frozen: (a) evolution of lead temperature in the cold leg; (b) evolution of steam SG output temperature; (c) evolution of SG pressure; and (d) evolution of reactor thermal power

Table 17 Safety margins estimation of the safety parameters of $T_{L,cold}$ sensor freezing cyber attack scenarios

Variable	T_{steam} (°C)	p_{SG} (Pa)	$T_{L,cold}$ (°C)	P_{Th} (W)
$\hat{y}^{\#1}_{\min,y_a}$	-	$1.7972 \cdot 10^7$	381.3042	$2.8412 \cdot 10^8$
$\hat{y}^{\#2}_{\max,y_a}$	528.1336	$1.8026 \cdot 10^7$	-	$3.0391 \cdot 10^8$
$M^{\#1(\text{or }\#2)}_{\#3}$	0.2187	0.9722	0.6260	0.4707

Note: 1) *a* in this Table refers to *$T_{L,cold}$ sensor freezing*, denoting that the simulation is run to simulate the system dynamic scenario processing when the $T_{L,cold}$ sensor is attacked to freezing and, to test the effects of such cyber attacks on the system integrity.

(2) Actuators

The results of the safety margins estimation of the three actuator failures are shown in Table 18. The cyber attacks leading to actuator-stuck failure at a random output level, severely affect the system functioning and integrity since most of the safety margins of the parameters turn out to be less than 0.2. This evidence should raise defenders' concern, because the ALFRED dynamics would be severely affected if cyber breaches are injected into these vulnerable components.

Table 18 Safety margins estimation of the safety parameters of the cyber attacks to actuators

Scenario *a*	$M_{U,T_{steam,a}}^{\gamma_1,\beta_1}$	$M_{T,p_{SG,a}}^{\gamma_1,\gamma_2,\beta}$	$M_{L,T_{L,cold,a}}^{\gamma_2,\beta_2}$	$M_{T,P_{Th,a}}^{\gamma_1,\gamma_2,\beta}$
CR height stuck	0.4895	0	0.7532	0
Water pump stuck	0	0	0.5441	0
Turbine valve stuck	0.1708	0	0.8484	0.1792

As illustrative example, Figure 26 shows the evolution of the safety parameters when the water pump is attacked to fail stuck with a random value sampled from the uniform distribution in $U(0{,}300)$ mentioned in Table 12, at a random time t_A. The feedwater mass flow rate G_{water} is output at a constant value in each case and this directly affects the SG performance. As a result, the lead temperature at the SG outlet $T_{L,cold}$ (Figure 26(a)) and steam SG outlet temperature T_{steam} (Figure 26(b)) are strongly affected. Then, changes in T_{steam} cause transients of SG pressure (Figure 26(c)), and, at the same time, $T_{L,cold}$ causes the CRs regulation that affects the reactor thermal power P_{Th} (Figure 26(d)). The results are shown in Table 19. Regarding the lower threshold, the safety margin of $T_{L,cold}$ turns out to be 0.5441, classified as a "medium" effect, according to three-level risk metric of Table 14. On the other hand, all safety margins of T_{steam}, p_{SG}, and P_{Th}, result to be equal to 0, indicating that a cyber attack to the water pump-stuck would "severely" affect the system dynamics and integrity.

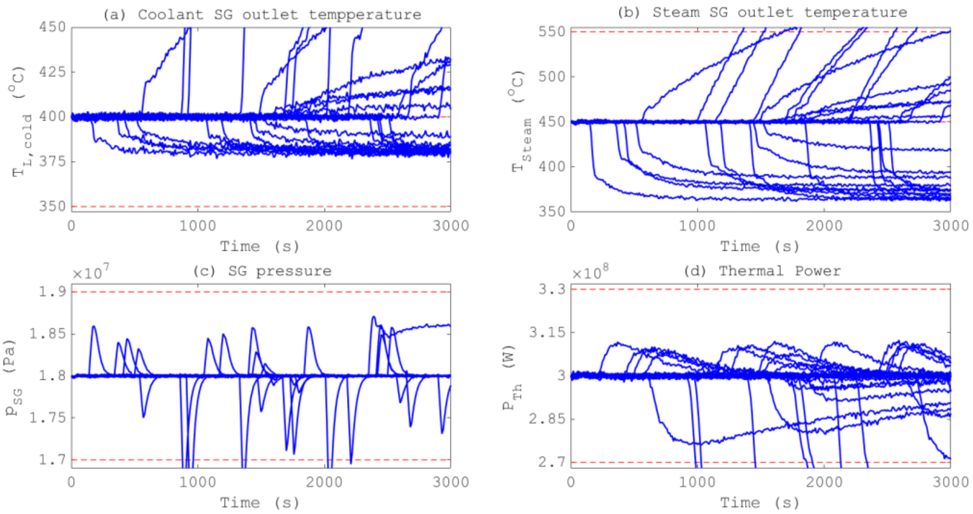


Figure 26 Profiles of the safety parameters for N_T=29 runs, when the water pump is attacked to fail stuck with a random value: (a) evolution of lead temperature in the cold leg; (b) evolution of steam SG output temperature; (c) evolution of SG pressure; and (d) evolution of thermal power

Table 19 Safety margins estimation of the safety parameters of water pump-stuck cyber attack scenarios

Variable	T_{steam} (°C)	p_{SG} (Pa)	$T_{L,cold}$ (°C)	P_{Th} (W)
$\hat{y}^{\#1}_{\min,y_a}$	-	$1.6089 \cdot 10^7$	377.2037	$2.0677 \cdot 10^8$
$\hat{y}^{\#2}_{\max,y_a}$	715.0707	$1.8712 \cdot 10^7$	-	$3.1242 \cdot 10^8$
$M^{\#1(\text{or } \#2)}_{\#3}$	0	0	0.5441	0

Note: 1) *a* in this Table refers to *water pump-stuck*, denoting that the simulation is run to test the system dynamic scenario processing when the water pump is attacked to get stuck in a random value and, to test the effects of such cyber attacks on the system integrity.

(3) PI controllers

The safety margins estimation results of cyber attacks to PI gains and set points are presented in Table 20. Cyber attacks to change of PI gain values have negligible effects on the safety parameters and on the system functionalities (except for changes of the K_p value of PI_3). This is potentially ascribed to the PI controller capability of regulating the errors of controlled variables close to zero even if the (relative small) gain values are changed to 3 or 4 orders of magnitude larger than the reference settings. On the other hand, cyber attacks changing the controllers set point values (i.e., $p_{SG,set}$, $T_{L,cold,set}$, $P_{Th,set}$) are more likely to cause system performance degradation. Such evidences demonstrate that PI gain values play a less important role, compared with the residual between the measurement and the set point value, $e(t)$.

Table 20 Safety margins estimation of the safety parameters of the cyber attacks to PI regulator value changes

Scenario a		$M_{U,T_{steam,a}}^{\gamma_1,\beta_1}$	$M_{T,p_{SG,a}}^{\gamma_1,\gamma_2,\beta}$	$M_{L,T_{L,cold,a}}^{\gamma_2,\beta_2}$	$M_{T,P_{Th,a}}^{\gamma_1,\gamma_2,\beta}$
PI_1	K_p	0.9676	0.9696	0.9203	0.9355
	K_i	0.9624	0.9698	0.9219	0.9232
	$T_{steam,set}$	0.9534	0.9591	0.9263	0.9295
PI_2	K_p	0.9612	0.9722	0.9304	0.9321
	K_i	0.9677	0.9684	0.9213	0.9370
	$p_{SG,set}$	0.9647	0.0213	0.9260	0.9300
PI_3	K_p	0.9451	0.7570	0.9156	0.8981
	K_i	0.9677	0.9660	0.9199	0.9414
	$T_{L,cold,set}$	0.6879	0.8840	0.5739	0.6264
PI_4	K_p	0.9623	0.9671	0.9168	0.9287
	K_i	0.9657	0.9699	0.9187	0.9343
	$P_{Th,set}$	0.9655	0.9685	0.9120	0.4628

Figure 27 shows the evolution of the safety parameters when the reference value of K_p of PI_1 is attacked at a random time t_A and changes to a random value distributed as U(1e-2,1) (see Table 13).

Under such circumstances, the steam SG outlet temperature T_{steam} (Figure 27(a)) is negligibly affected. The most probable reason is that K_p plays a less important role in PI computation, compared with the residual between the measurement T_{steam} and the set point value $T_{steam,set}$, $e(t)$. The resulting negligible change of T_{steam} will not lead to any transients of SG functioning. Also, the evolutions of SG pressure p_{SG} (Figure 27(b)), of lead temperature at the SG outlet $T_{L,cold}$ (Figure 27(c)), and of reactor thermal power P_{Th} (Figure 27(d)) are not altered with respect to normal conditions. Safety margins of T_{steam}, p_{SG}, $T_{L,cold}$ and P_{Th} result to be equal to 0.9676, 0.9696, 0.9203, and 0.9355, respectively, as listed in Table 21.

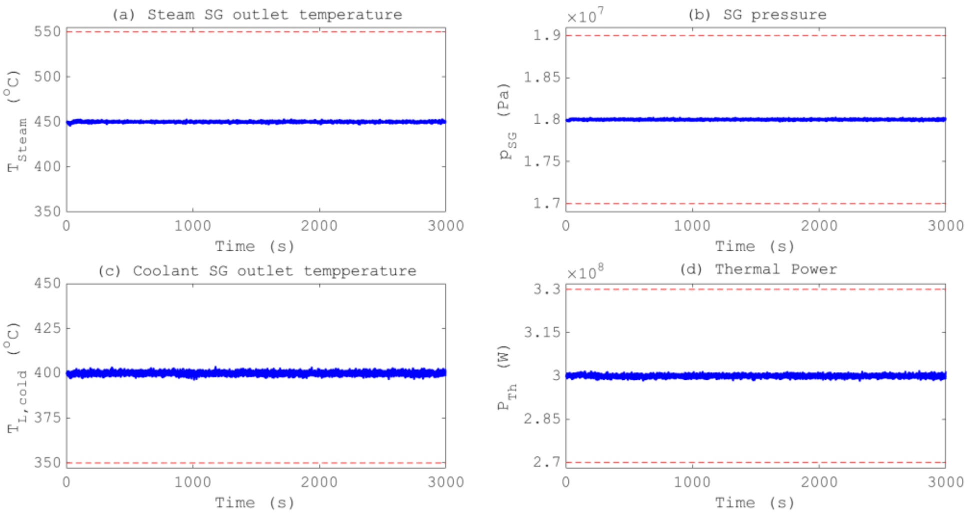


Figure 27 Profiles of the safety parameters for N_T=29 runs, when the PI_1 Kp gain is changed to a random value: (a) evolution of steam SG output temperature; (b) evolution of SG pressure; (c) evolution of lead temperature in the cold leg; and (d) evolution of reactor thermal power

Table 21 Safety margins estimation of the safety parameters of change of K_p value of PI_1 cyber attack scenarios

Variable	T_{steam} (°C)	p_{SG} (Pa)	$T_{L,cold}$ (°C)	P_{Th} (W)
$\hat{y}^{\#1}_{\min,y_a}$	-	$1.7971 \cdot 10^7$	396.0125	$2.9809 \cdot 10^8$
$\hat{y}^{\#2}_{\max,y_a}$	453.2381	$1.8030 \cdot 10^7$	-	$3.0193 \cdot 10^8$
$M^{\#1(\text{or } \#2)}_{\#3}$	0.9676	0.9696	0.9203	0.9355

Note: 1) *a* in this Table refers to *change of K_p value of PI_1*, denoting that the simulation is run to test the system dynamic scenario processing when the K_p gain value of PI_1 is attacked to be changed to a random value and, to test the effects of such cyber attacks on the system integrity.

3.4. Conclusions

Tasks I.2, II.2 and III.2 of Table 1 have been the focus of the Ph.D. activities on the analysis of threats to CPSs. The application of the approach has been illustrated on the digital I&C system of an Advanced Lead-cooled Fast Reactor European Demonstrator (ALFRED).

A modelling and simulation framework has been developed for identifying components vulnerable to cyber threats in CPSs of NPPs, and for generating and processing cyber attack scenarios in CPSs accounting for multiple failure modes of the attacked components, to test the effects of the cyber threats on the system functionality and integrity, and to prioritize the most vulnerable components for cyber security protection decision-making. Cyber attacks are injected by a MC engine of cyber breaches injection. A safety margin estimation approach has been utilized for cyber threat prioritization in Section 3.2. Safety margins of the safety parameters are estimated by a Bracketing Order Statistics approach, with respect to the one- and two-sided thresholds. The results of Section 3.3 obtained from the case study of the ALFRED identify actuators as the most vulnerable CPS components. The cyber attacks leading to actuator-stuck failure at a random output level severely affect the system functioning since most of the safety margins of the parameters turn out to be less than the threshold of 0.2 denoting high risk. This evidence should raise defenders' concern, because the ALFRED dynamics would be severely affected if cyber breaches are injected into these vulnerable components (Paper 3[J]).

4. MODELING AND SIMULATION FOR PROTECTION DESIGN

Contents of the Chapter have been adapted from:

4[J] *Wang, W., Di Maio, F. and Zio, E., 2018. Adversarial Risk Analysis to Allocate Optimal Defense Resources for Protecting Nuclear Power Plants from Cyber Attacks, Risk Analysis, under review.*

1[B] *Wang, W., Di Maio, F. and Zio, E., 2018. A Non-Parametric Cumulative Sum Approach for Real-Time Diagnostics of Cyber Attacks to Nuclear Power Plants. Resilience of Cyber-Physical Systems: From Risk Modelling to Threat Counteraction, Chapter 9. DOI: 10.1007/978-3-319-95597-1.*

5[J] *Wang, W., Di Maio, F. and Zio, E., 2018. Reliability Assessment of an Online Cyber Security Diagnostic Tool of a Nuclear Power Plant under Uncertain Human Operator Cognition. Work in progress.*

For modeling and simulation for CPSs protection design, an Adversarial Risk Analysis (ARA) approach is developed in Chapter 4.1 for obtaining a novel prescriptive defender support strategy that optimize both the allocation of resources for defensive barriers against cyber attacks and the maintenance strategy to cope with component stochastic failures; and for online diagnostics of cyber attacks to CPSs, a Non-Parametric Cumulative Sum (NP-CUSUM) detection approach is developed in Chapter 4.2, to promptly recognize cyber attacks, distinguish them from component failures, and guiding decisions for CPSs recovery.

4.1. Defend-attack modeling for optimal allocation of resources for defensive barriers

The intelligent and adaptive nature of attackers, that with intentional and malicious attacks aim at maximizing vulnerable components loss, endangers the CPS functionality; adversarially, defenders have to enforce defense strategies by taking decisions regarding resource allocations, to protect the integrity and survivability of CPSs from intentional and malicious cyber threats. The minimization of attacks impacts on CPS functionality and the maximization of CPS reliability and survivability are sought by defenders decisions on the allocation of defensive resources.

In this Chapter, we propose an ARA approach to provide a novel one-sided prescriptive support strategy for the defender to optimize the defensive resource allocation, based on a subjective expected utility model, in which the decisions of the adversaries are uncertain. This increases confidence in cyber security through robustness of CPS protection actions against uncertain malicious threats, compared with prescriptions provided by a classical defend-attack game-theoretical approach.

4.1.1. Case study: the ALFRED

We consider the protection of the ALFRED against potential cyber threats. Cyber attacks to the ALFRED and the deployable defensive resources are presented in Sections 4.1.1.1 and 4.1.1.2, respectively.

4.1.1.1. The cyber attacks

Besides components failures, CPS functionality can also be compromised by malicious attacks. Responses of the digital I&C system of ALFRED to 15 different cyber attack strategies aimed at altering sensors, actuators and PI regulators (i.e., PI gains and set point values) have been investigated in Chapter 3. It is shown that cyber attacks to actuators challenge the most the entire system functionality, along with the attacks to the lead temperature sensor, whereas, functionality is negligibly affected by attacks that alter the values of PI gains. This is ascribed to the PI controller capability of regulating the errors of controlled variables close to zero even if the (relatively small) gain values are changed to 3 or 4 orders of magnitude larger than the reference settings. It is worth pointing out that the prioritization of cyber threats in terms of their impact on the ALFRED functionality, as proposed in Chapter 3, are usually unknown (i.e., uncertain) to attackers, or not equally perceived by attackers and defenders, at least in reality.

In this Chapter, a poor attacker cognitive awareness on cyber threats prioritization is assumed, resulting in a pool of A=15 different cyber attack strategies (of 4 types, as listed in Table 22, namely, a_1 (attacks to different sensor databases); a_2 (attacks to commands of different actuators); a_3 (attacks to changes of PI gain values); a_4 (attacks to changes of set point values of controlled variables)) that the attacker can undergo, constrained by resources that allow him/her to launch a single attack to target a single CPS component.

An intentional attack can be launched either from an outsider or from an insider (e.g., a bribed operator) with probabilities ξ_{out} and ξ_{in} (hereafter taken equal to $\xi_{out} = 0.99$ and $\xi_{in} = 0.01$, respectively) with preparation cost [206, 207]:

$$c_{prep} = \begin{cases} c_{out}, & \text{if outsider attacker} \\ c_{in}, & \text{if insider attacker} \end{cases} \tag{4-1}$$

where, c_{out} and c_{in} are the front money for financing an outsider attacker and the bribery cost of an insider operator (hereafter assumed to be distributed as truncated normal distributions TN(5e1, 1e1) (k€) and TN(2e2, 5e1) (k€), respectively, according to the statistics listed in [208]).

Attack consequences can be monetized in terms of attack loss due to c_{A1} (attacker arrest) and c_{A2} (cyber attacker remunerations of launching an attack), and attack revenues from c_{A3} (radiological effect), c_{A4} (public panic effect) and c_{A5} (media effect). The total attack cost becomes:

$$c_A = c_{prep} + \sum_{q=1}^{2} c_{Aq} - \sum_{q=3}^{5} c_{Aq} \tag{4-2}$$

Notice that:

(1) Loss c_{A1} (i.e., cost for an attacker arrest by a security personnel (e.g., police office)) is here estimated by:

$$c_{A1} = \xi_{arrest} \cdot c_{arrest} \tag{4-3}$$

where, ξ_{arrest} is the arrest probability hereafter assumed to be distributed as a Uniform distribution U(0.0, 0.5), and c_{arrest} is equal to 3e2 (k€) [168, 209, 210].

(2) The attacker remunerations c_{A2} are usually deliberated between the attacker and the employer before launching an attack, according to uncertain factors such as attacker experience, attack technical means, etc.; thus, c_{A2} is estimated to be several times larger than the front money, hereby distributed as a truncated normal distribution TN(1e3, 2e2) (k€).

(3) c_{A3}, c_{A4} and c_{A5} are the attacker revenues induced from the launched attack and mainly depend on the confrontation between the attack and any possible defensive countermeasures.

4.1.1.2. The defensive resources

A typical digital I&C system is a SCADA (Supervisory Control And Data Acquisition) system that features numerous hardwares and softwares, interfacing the monitoring and control system with the physical process, aimed at controlling it and, at the same time, protecting it from cyber attacks, from which recovery is needed (in case of attack success) to maintain the system in normal operation conditions [101, 211].

Defensive resources are, therefore, aimed at: (d_1) preventing from cyber attacks and, (d_2) recovering when suffering a successful cyber attack.

Prevention can be enforced by [53, 101, 212, 213]:

- Firewall that prevents intrusions and blocks unauthorized or unwanted communications;
- Intrusion Detection Systems (IDSs) that identify common patterns of unwanted network access or malicious activities, and alert operators;
- Operators that monitor the process status through sophisticated Human-Machine Interfaces (HMIs) that embed IDSs distinguishing cyber attacks from stochastic component failures;

- Security software that prevents from operators unauthorized access and information leakage by password authentication, communication encryption, or/and access authorization;

Recovery from successful cyber attacks can rely on [212]:

- Mainframe computers that allow the digital I&C system to run interrupted and provide correct commands to actuators, even under some types of cyber attacks (such as Internet Protocol (IP) spoofing);
- Database servers that store clean databases and can be used for recovery of data in case of some types of cyber attacks (such as false data injection);
- Security engineers that maintain the digital I&C system once exposed to cyber attack, to guarantee a secure network communication and service.

Table 23 lists the defensive resources considered, with their relevance to cope with the cyber attacks discussed in Section 4.1.1.1 (Column 3) and their minimum and maximum deployable quantity (Column 5), both assessed by expert judgment [101, 211, 213].

The annual costs of deployment of defensive resources (Column 6) are estimated on salaries (for operators and security engineers), software research and development (R&D) (for firewall and security software), and equipment costs (for IDSs, mainframe computers and database servers). In details, salaries correspond to annual base wages and pay incentives, whereas costs of software R&D and equipment are estimated as in Eqs. (4-4) and (4-5), respectively [214]:

$$c_{i,k} = \frac{c_{i,k}^{R\&D}}{T_{NPP}} + c_{i,k}^{M}\text{, if } c_{i,k} = c_{1,1},\ c_{1,4} \tag{4-4}$$

$$c_{i,k} = \frac{c_{i,k}^{Buy}}{T_{equipment}} + c_{i,k}^{E}\text{, if } c_{i,k} = c_{1,2},\ c_{2,1},\ c_{2,2} \tag{4-5}$$

where, $c_{i,k}$ is the annual cost of the *k*-th resource of the *i*-th defense type, $c_{i,k}^{R\&D}$ is the R&D cost, T_{NPP} is the lifetime of a ALFRED NPP, $c_{i,k}^{M}$ is its annual maintenance cost, $c_{i,k}^{Buy}$ is its purchase cost, amortized for its lifetime $T_{equipment}$ (without depreciation), and $c_{i,k}^{E}$ is the annual cost of electricity needed to run the resource.

Table 22 Cyber attack strategies: types and targets

Attack type (a_j)	Probability of attack success (probability of prevention failure), $\phi_{s_1}^j$	Probability of recovery failure (if attack success), $\phi_{s_2}^j$	Attack target, ($a_{j,y}$)			
(a_1) sensor databases	0.65	0.40	($a_{1,1}$) T_{steam}	($a_{1,2}$) p_{SG}	($a_{1,3}$) $T_{L,cold}$	($a_{1,4}$) P_{Th}
(a_2) commands of actuators	0.55	0.45	($a_{2,1}$) h_{CR}	($a_{2,2}$) G_{water}	($a_{2,3}$) kv	/
(a_3) changes of PI gain values	0.40	0.80	($a_{3,1}$) PI_1	($a_{3,2}$) PI_2	($a_{3,3}$) PI_3	($a_{3,4}$) PI_4
(a_4) changes of set point values	0.40	0.50	($a_{4,1}$) $T_{steam,set}$	($a_{4,2}$) $p_{SG,set}$	($a_{4,3}$) $T_{L,cold,set}$	($a_{4,4}$) $P_{Th,set}$

Table 23 Defensive resources with properties

Defense type, d_i	Countermeasures, $x_{i,k}$	Relevance	$x_{i,k}$ relevance with respect to (a_j), $\gamma_{i,k}^j$				Min.-Max., $n^r_{i,k}$	Annual cost distribution, $c_{i,k}$ (k€)
			(a_1)	(a_2)	(a_3)	(a_4)		
(d_1) Prevention	($x_{1,1}$) firewall	High	0.25	0.25	0.25	0.25	0-1	TN(80,20)
	($x_{1,2}$) Intrusion Detection Systems (IDSs)	Moderate	0.10	0.10	0.01	0.05	0-4	TN(15,2)
	($x_{1,3}$) operators	Moderate	0.10	0.10	0.01	0.05	1-4	Tri(35,50,60)
	($x_{1,4}$) security software	Moderate	0.06	0.06	0.10	0.10	0-3	TN(80,10)
(d_2) Recovery	($x_{2,1}$) mainframe computers	High	0.17	0.45	0.35	0.35	0-3	TN(520,2)
	($x_{2,2}$) database servers	High	0.25	0.25	0.05	0.15	0-2	TN(70,2)
	($x_{2,3}$) security engineers	High	0.50	0.35	0.25	0.25	0-2	Tri(90,100,110)

Notes: Tri(a,b,c) denotes a triangular distribution with lower limit a, upper limit c and mode b, and TN(μ,σ) denotes a normal distribution with mean value μ and standard deviation σ, truncated at zero.

Considering a maximum budget B_M generates a set of $\Re$ alternative defense portfolios $d^r = \{d_1^r, d_2^r\} = \{n_{1,1}^r, n_{1,2}^r, n_{1,3}^r, n_{1,4}^r, n_{2,1}^r, n_{2,2}^r, n_{2,3}^r\} \in \Re$ characterized by an annual cost c_{annual}^r, r = 1, 2, …, $\Re$:

$$c_{annual}^{r} = c_{annual}^{d_1^r} + c_{annual}^{d_2^r} = \sum_i \sum_k n_{i,k}^r \cdot c_{i,k} \leq B_M \tag{4-6}$$

where, $c_{annual}^{d_1^r}$ and $c_{annual}^{d_2^r}$ are the annual costs of the d_1 and d_2 types of defensive resources of the portfolio d^r. Assuming a B_M equal to 2,000 k€ (for sake of illustration), Eq. (4-6) yields $\Re$ = 4834 alternative defensive resource allocations with $n_{1,1}^r = 0{,}1$, $n_{1,2}^r = 0{,}1{,}2{,}3{,}4$, $n_{1,3}^r$=1,2,3,4, $n_{1,4}^r$=0,1,2,3, $n_{2,1}^r$=0,1,2,3, $n_{2,2}^r$=0,1,2 and $n_{2,3}^r$=0,1,2, and $c_{i,k}$ are taken to be the mean values of the annual costs of the defensive resources. The resulting $\Re$ = 4834 deployable portfolios are hereafter referred to by the rule of sequentially increasing the values of $n_{2,3}^r, n_{2,2}^r, n_{2,1}^r, n_{1,4}^r, n_{1,1}^r, n_{1,2}^r$ and $n_{1,1}^r$, and, thus, lead to the permutations $d^1 = \{0,0,1,0,0,0,0\}$, $d^2 = \{0,0,1,0,0,0,1\}$, …, and $d^{4834} = \{1,4,4,3,2,2,2\}$.

4.1.2. The defend-attack model

The minimization of attacks impact on CPS functionality and the maximization of CPS reliability and survivability are sought by defenders decisions on the allocation of defensive resources [117, 215-218]. A variety of defend-attack models have been proposed for this scope, focusing on the strategic interactions between defenders and attackers for the optimal defense resource allocation. Often, game-theoretical models (widely applied to many areas such as economics, political science, psychology, biology, computer science, and so on [114-116]) have been used for advising the defender on the optimal allocation of defensive resources against attackers [40, 80, 100, 117, 118, 219].

However, all models mentioned above are developed from the viewpoint of a neutral opponent governing the attack/defense loss, under the strong assumptions of mutually common knowledge (i.e., Nash equilibrium), rather than from the viewpoint of an intelligent adversary (attacker or defender) exploring the impacts of malicious (or self-interested) actions under uncertainty [104-107]. Adversarial Risk Analysis (ARA) addresses this limitation by modeling and analyzing intelligent actors (attackers or defenders), for which the outcomes (or losses) in the game-theoretical model are uncertain [104, 106, 169].

In this work, we propose an ARA model to advise the CPS defender, with his own beliefs and preferences, for identifying the optimal defense resource allocation that would minimize the system integrity loss when constrained by limited defense resources against (unknown and uncertain) cyber attacks. The game originated between the defender and the attacker is described in Section 4.1.2.1. An ARA model [104, 106, 169] is here tailored to the problem of cyber security assessment of the digital I&C system of the ALFRED, for supporting the defender to allocate the optimal defenses under uncertain adversarial strategies and consequences of attacks. The resulting ARA model is also compared with a Nash equilibrium optimal solution of a classical game-theoretical analysis [220-222], where uncertainties are neglected.

4.1.2.1. The game

The defender of the ALFRED needs to choose a defense strategy d^r from the $\mathfrak{R}$ available, to optimally protect the digital I&C system from an (unknown) attack strategy $a_{j,k}$ among the A that can threaten the system, originating a game between the defender and the attacker. Different combinations of defense and attack strategies $\left(\vec{d}^r, a_{j,y}\right) = \left(d_1^r, d_2^r, a_{j,y}\right)$ would result in different outcomes and consequences, with different costs for both the defender and the attacker.

Since the scope of the work is to prescriptively support the defender with an optimal resource allocation, outcomes and consequences generated from each $\left(d_1^r, d_2^r, a_{j,y}\right)$ are hereafter described only with focus on the defender decision making.

4.1.2.1.1. The outcomes probabilities

Each combination $\left(d_1^r, d_2^r, a_{j,y}\right)$ originates the outcome set $\vec{s} = \left\{s_1\left(d_1^r, a_{j,y}\right), s_2(d_2^r|s_1)\right\}$, where $s_1\left(d_1^r, a_{j,y}\right)$ defines the successful prevention of d_1^r to an attack $a_{j,y}$:

$$s_1\left(d_1^r, a_{j,y}\right) = \begin{cases} 1, & \text{prevention failure (attack success)} \\ 0, & \text{prevention success (attack failure)} \end{cases} \tag{4-7}$$

and $s_2(d_2^r|s_1)$ the successful recovery of d_2^r in case of successful attack (i.e., $s_1(d_1^r, a_{j,y})$=1):

$$s_2(d_2^r|s_1) = \begin{cases} 1, & \text{recovery success} \\ 0, & \text{recovery failure} \end{cases} \tag{4-8}$$

The outcome set comes with a probability set $\vec{p} = \left\{p\left(s_1(d_1^r, a_{j,y})\right), p\left(s_2(d_2^r|s_1)\right)\right\}$, where $p\left(s_1(d_1^r, a_{j,y})\right)$ defines the probability of the prevention outcome $s_1(d_1^r, a_{j,y})$, and $p\left(s_2(d_2^r|s_1)\right)$ the probability of the recovery outcome $s_2(d_2^r|s_1)$ in case of successful attack. As proposed in [168], the values of $p\left(s_1(d_1^r, a_{j,y})\right)$ and $p\left(s_2(d_2^r|s_1)\right)$ are calculated as in Eqs. (4-9) and (4-10), respectively,

$$p\left(s_1\left(d_1^r, a_{j,y}\right)\right) = \begin{cases} \phi_{s_1}^j \cdot \exp(-\sum_k n_{1,k}^r \cdot \gamma_{1,k}^j); & s_1\left(d_1^r, a_{j,y}\right) = 1 \\ 1 - \phi_{s_1}^j \cdot \exp(-\sum_k n_{1,k}^r \cdot \gamma_{1,k}^j); & s_1\left(d_1^r, a_{j,y}\right) = 0 \end{cases} \tag{4-9}$$

$$p\left(s_2\left(d_2^r \mid s_1\right)\right) = \begin{cases} 1 - \phi_{s_2}^j \cdot \exp(-\sum_k n_{2,k}^r \cdot \gamma_{2,k}^j); & s_2\left(d_2^r \mid s_1\right) = 1,\ s_1\left(d_1^r, a_{j,y}\right) = 1 \\ \phi_{s_2}^j \cdot \exp(-\sum_k n_{2,k}^r \cdot \gamma_{2,k}^j); & s_2\left(d_2^r \mid s_1\right) = 0,\ s_1\left(d_1^r, a_{j,y}\right) = 1 \\ 1; & s_1\left(d_1^r, a_{j,y}\right) = 0 \end{cases} \tag{4-10}$$

where, $\phi_{s_1}^j$ is the probability of prevention failure when the *j*-th type attack is occurring, $\phi_{s_2}^j$ is the probability of recovery failure when the *j*-th type attack is successful (see Table 22), $\gamma_{1,k}^j$ is the estimated relevance of the *k*-th prevention countermeasure in decreasing the attack success probability, and $\gamma_{2,k}^j$ is the estimated relevance parameter of the *k*-th recovery measure in increasing the recovery success probability (see Table 23).

Being all these parameters estimated by the defender on his personal judgment, Eqs. (4-9) and (4-10) are the defender opinion on the outcomes probabilities, i.e., $p_D\left(s_1(d_1^r, a_{j,y})\right)$ and $p_D\left(s_2(d_2^r|s_1)\right)$. However, the defender ignores the attacker assumptions on the probabilities of the outcomes, i.e., $p_A\left(s_1(d_1^r, a_{j,y})\right)$ and $p_A\left(s_2(d_2^r|s_1)\right)$, and can only speculate assuming them to be distributed as normal distributions with $p_D\left(s_1(d_1^r, a_{j,y})\right)$ and $p_D\left(s_2(d_2^r|s_1)\right)$ as mean values, and Eqs. (4-11) and (4-12) as standard deviations [168],

$$\sigma_A\left(s_2 \mid d_2^r, s_1\right) = \min(p_D\left(s_2 \mid d_2^r, s_1\right), 0.05) \tag{4-11}$$

$$\sigma_A\left(s_1 \mid d_1^r, a_{j,y}\right) = \min(p_D\left(s_1 \mid d_1^r, a_{j,y}\right), 0.05) \tag{4-12}$$

4.1.2.1.2. The attack consequences

Consequences of attacks are monetized in terms of economic loss (i.e., for c_{D1} system integrity loss and c_{D2} decrease of P_{Mech}) and compensation for post-attack impact (i.e., c_{D3} radiological effects, c_{D4} public panic and chaos, and c_{D5} media impact) (see Table 24) [223, 224].

Table 24 Consequences of attacks

Consequences	Description
(c_{D1}) System integrity loss	CPS recovery and protection improvement
(c_{D2}) Decrease of P_{Mech}	Business interruption
(c_{D3}) Radiological effects	Compensation for radiation pollution
(c_{D4}) Public panic and chaos	Social network reconstruction
(c_{D5}) Media impact	Public relation management

For simplicity, and in line with [84, 225], the costs of the l-th consequence $c_{Dl}\left(\vec{s}\middle|\vec{d}^r, a_{j,y}\right)$, l=1, 2, …, 5, that depend on $\left(\vec{d}^r, a_{j,y}\right)$ and on the outcomes $\vec{s}$, is calculated according to the law of total probability [226]:

$$c_{Dl}\left(\vec{s} \mid d^r, a_{j,y}\right) = \sum_{\alpha} p_l^{\alpha}\left(\vec{s} \mid a_{j,y}\right) \cdot c_{Dl}^{\alpha} \tag{4-13}$$

where, c_{Dl}^{α} is assumed to be the cost of a negligible (α=N), medium (α=M) and severe (α=S) attack (listed in Table 25), $p_l^{\alpha}\left(\vec{s}\middle|a_{j,y}\right)$ is the conditional probability of the outcome $\vec{s}$ to the occurrence of $a_{j,y}$ with the α-th effect of the l-th consequence, and:

$$\sum_{\alpha} p_l^{\alpha}\left(\vec{s} \mid a_{j,y}\right) = 1 \tag{4-14}$$

Generally, the defender empirically assesses $p_l^{\alpha}\left(\vec{s}\middle|a_{j,y}\right)$ taking three rules into account:

(1) α: besides the consideration of Eq. (4-14), since the game considers the cyber attack to a single element, the defender empirically assumes that an outcome $\vec{s}$ (i.e., $(s_2 = 1, s_1 = 1)$, $(s_2 = 0, s_1 = 1)$ or $(s_1 = 0)$) of a cyber attack $a_{j,y}$ leads either to a negligible (α=N) or medium (α=M) l-th consequence rather than to a severe (α=S) effect. For example, given an outcome $(s_2 = 1, s_1 = 1)$ of $a_{1,2}$, the values of $p_1^N\left(s_2 = 1, s_1 = 1\middle|a_{1,2}\right)$ (equal

to 0.80) and $p_1^M(s_2 = 1, s_1 = 1|a_{1,2})$ (equal to 0.17) turn out to be much larger than $p_1^S(s_2 = 1, s_1 = 1|a_{1,2})$ (equal to 0.03) in Table 27(I);

(2) $(\vec{s}|a_{j,y})$: in general terms, the failure recovery in case of a successful attack $a_{j,y}$, i.e., $(s_2 = 0, s_1 = 1)$ is more likely to lead the l-th consequence to a severe (α=S) effect, whereas, the successful prevention to $a_{j,y}$ $(s_1 = 0)$ more probably has a negligible (α=N) effect on the l-th consequence. For example, given the cyber attack $a_{1,2}$, the defender assumes the relatively large values of $p_1^M(s_2 = 0, s_1 = 1|a_{1,2})$ (equal to 0.40) and $p_1^S(s_2 = 0, s_1 = 1|a_{1,2})$ (equal to 0.05) but a small value of $p_1^S(s_1 = 0|a_{1,2})$ equal to 0.00 in Table 27(I);

(3) l: the responses of ALFRED to cyber attacks to sensors, actuators and PI regulators are investigated by simulation in Chapter 3 to quantify the effects of the cyber attacks on the system functionality and integrity. For this scope, we rely on the safety margins estimates of the ALFRED under different cyber attacks (listed in Tables 16, 18 and 20 in Chapter 3). In general terms, the smaller the safety margin to cyber attack, the more probable a l-th consequence with severe (α=S) effect (for example, the failure recovery in case of a successful attack to water pump (actuator) $a_{2,2}$, i.e., $p_1^S(s_2 = 0, s_1 = 1|a_{2,2})$ (equal to 0.20) is more easily to lead to the severe (α=S) effect on the system integrity loss than a successful attack to T_{steam} sensor $a_{1,1}$, i.e., $p_1^S(s_2 = 0, s_1 = 1|a_{1,1})$ (equal to 0.00) in Table 27(I), because Table 20 shows cyber attacks leading to actuator-stuck failure at a random output level severely affect the system functioning and leads most of the safety margins of the parameters to turn out to be less than the threshold of 0.2 denoting high risk). Besides, it is also assumed that, once prevented $(s_1 = 0)$, the attack has no opportunity to lead the system to a l-th consequence with severe (α=S) effect, hence $p_1^S(s_1 = 0|a_{1,1})$ turns out to be equal to 0.00, for example, in Table 27(I).

In Table 27, the defender assumptions for $p_l^\alpha(\vec{s}|a_{j,y})$ (for system integrity loss (I), decrease of P_{Mech} (II), radiological effects (III), public panic and chaos (IV) and media impact (V)) are listed.

Table 25 Defender's assessment of base costs of the consequences

Consequences	Cost of each level (k€)		
	Negligible	Medium	Severe
	c_{Dl}^{N}	c_{Dl}^{M}	c_{Dl}^{S}
(c_{D1}) System integrity loss	1e1	4e2	1e3
(c_{D2}) Decrease of P_{Mech}	91.11	91.11*24	91.11*1e2
(c_{D3}) Radiological effect	0	1e4	1e6
(c_{D4}) Public panic and chaos	1e1	1e3	1e4
(c_{D5}) Media impact	1e1	1e3	1e4

Notes: 91.11 (k€/h) is an estimate of the economics profit per hour for a 300MW nuclear reactor.

In conclusion, with respect to a generic $(\vec{d}^r, a_{j,y})$ the total cost to be considered for decision-making consists in the defenses deployment cost c_{annual}^{r} of Eq. (6) plus the sum of all possible consequences costs $c_{Dl}(\vec{s}|d^r, a_{j,y})$, l=1, 2, 3, 4, 5:

$$c_D\left(\vec{s} \mid d^r, a_{j,y}\right) = c_{annual}^{r} + \sum_{l=1}^{5} c_{Dl}\left(\vec{s} \mid d^r, a_{j,y}\right) \tag{4-15}$$

The attacker must sustain the attack impact (i.e., c_{D3}, c_{D4} and c_{D5}), that has to be justified in light of the speculated costs resulting from the game, i.e., c_{A3}, c_{A4} and c_{A5}, by:

$$c_{Aq}\left(\vec{s} \mid d^r, a_{j,y}\right) = \sum_{\alpha} p_q^{\alpha}\left(\vec{s} \mid a_{j,y}\right) \cdot c_{Aq}^{\alpha}, \quad q = 3,4,5 \tag{4-16}$$

where c_{Aq}^{α} is the cost of a negligible (α=N), medium (α=M) and severe (α=S) attack (see Table 26 their personal distributions) and $p_q^{\alpha}(\vec{s}|a_{j,y})$ is the conditional probability of the outcome $\vec{s}$ to the occurrence of $a_{j,y}$ with the α-th effect of the q-th consequence. Thus, the attacker costs of Eq. (4-2) become:

$$c_A\left(\vec{s} \mid d^r, a_{j,y}\right) = c_{prep} + c_{A1} + c_{A2} + \sum_{q=3}^{5} c_{Aq}\left(\vec{s} \mid d^r, a_{j,y}\right) \tag{4-17}$$

Table 26 Assessment of the distributions of base costs of the consequences

Consequences (q)	Cost of each level c_{Aq}^{α} (k€)		
	Negligible	Medium	Severe
	c_{Aq}^{N}	c_{Aq}^{M}	c_{Aq}^{S}
(iii) Radiological effect	0	TN(1e4,5e3)	TN(1e6,5e5)
(iv) Panic effect	0	TN(1e3,5e2)	TN(1e4,5e3)
(v) Media effect	0	TN(1e4,5e3)	TN(1e5,5e4)

In decision analysis, it is common to map the cost into a utility function that measures the decision maker preference on alternatives with uncertain outcomes [227, 228]. The decision maker aims at optimizing his/her portfolio by maximizing his/her own utility function [229, 230].

In our case, the defender may use the exponential utility $u_D(\vec{s}|d^r, a_{j,y})$ of Eq. (4-18), that is a risk averse function that lowers the uncertainty of consequences (i.e., costs) assuming constant absolute risk with the coefficient of risk aversion $\Lambda_D = -u''_D(\vec{s}|d^r, a_{j,y})/u'_D(\vec{s}|d^r, a_{j,y}) < 0$ [104, 231]:

$$u_D\left(\vec{s} \mid d^r, a_{j,y}\right) = -\exp\left(k_D \cdot c_D\left(\vec{s} \mid d^r, a_{j,y}\right)\right) \tag{4-18}$$

where, k_D (here taken distributed as U(1e-5, 2e-5)) is defined according to $k_D = u''_D(\vec{s}|d^r, a_{j,y})/u'_D(\vec{s}|d^r, a_{j,y}) > 0$, whose absolute value is constant with respect to costs and larger than 0 [232, 233].

Relying on the concept of exponential utility with attacker's risk proneness attitude and coefficient of risk proneness $\Lambda_A = -u''_A(\vec{s}|d^r, a_{j,y})/u'_A(\vec{s}|d^r, a_{j,y}) > 0$ [104, 231-233], the defender can assess the attacker's utility $u_A(\vec{s}|d^r, a_{j,y})$ as:

$$u_A\left(\vec{s} \mid d^r, a_{j,y}\right) = \exp\left(k_A \cdot c_A\left(\vec{s} \mid d^r, a_{j,y}\right)\right) \tag{4-19}$$

where, k_A is estimated based on the absolute risk proneness constant with respect to costs and $k_A = u''_A(\vec{s}|d^r, a_{j,y})/u'_A(\vec{s}|d^r, a_{j,y}) < 0$, judged to be distributed from a uniform distribution U(-3.0e-5, 0) with the aid of experts.

Table 27 Defender assessment of the probabilities of occurrence of each consequence level

(I) System integrity loss

Probabilities	Attack strategies														
	$a_{1,1}$	$a_{1,2}$	$a_{1,3}$	$a_{1,4}$	$a_{2,1}$	$a_{2,2}$	$a_{2,3}$	$a_{3,1}$	$a_{3,2}$	$a_{3,3}$	$a_{3,4}$	$a_{4,1}$	$a_{4,2}$	$a_{4,3}$	$a_{4,4}$
$p_1^N(s_2 = 0, s_1 = 1 \| a_{j,y})$	1.00	0.55	0.40	0.35	0.50	0.30	0.50	1.00	1.00	0.90	1.00	1.00	0.65	0.40	0.65
$p_1^M(s_2 = 0, s_1 = 1 \| a_{j,y})$	0.00	0.40	0.50	0.55	0.50	0.50	0.50	0.00	0.00	0.10	0.00	0.00	0.35	0.50	0.35
$p_1^S(s_2 = 0, s_1 = 1 \| a_{j,y})$	0.00	0.05	0.10	0.10	0.00	0.20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.00
$p_1^N(s_2 = 1, s_1 = 1 \| a_{j,y})$	1.00	0.80	0.75	0.70	0.75	0.65	0.75	1.00	1.00	1.00	1.00	1.00	0.85	0.80	0.85
$p_1^M(s_2 = 1, s_1 = 1 \| a_{j,y})$	0.00	0.17	0.20	0.25	0.20	0.30	0.22	0.00	0.00	0.00	0.00	0.00	0.15	0.20	0.15
$p_1^S(s_2 = 1, s_1 = 1 \| a_{j,y})$	0.00	0.03	0.05	0.05	0.05	0.05	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$p_1^N(s_1 = 0 \| a_{j,y})$	1.00	0.99	0.99	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.99
$p_1^M(s_1 = 0 \| a_{j,y})$	0.00	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.01
$p_1^S(s_1 = 0 \| a_{j,y})$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

(II) Decrease of PMech

Probabilities	Attack strategies														
	$a_{1,1}$	$a_{1,2}$	$a_{1,3}$	$a_{1,4}$	$a_{2,1}$	$a_{2,2}$	$a_{2,3}$	$a_{3,1}$	$a_{3,2}$	$a_{3,3}$	$a_{3,4}$	$a_{4,1}$	$a_{4,2}$	$a_{4,3}$	$a_{4,4}$
$p_2^N(s_2 = 0, s_1 = 1 \| a_{j,y})$	1.00	0.55	0.40	0.35	0.50	0.30	0.50	1.00	1.00	0.90	1.00	1.00	0.65	0.40	0.65
$p_2^M(s_2 = 0, s_1 = 1 \| a_{j,y})$	0.00	0.40	0.50	0.55	0.50	0.50	0.50	0.00	0.00	0.10	0.00	0.00	0.35	0.50	0.35
$p_2^S(s_2 = 0, s_1 = 1 \| a_{j,y})$	0.00	0.05	0.10	0.10	0.00	0.20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.00
$p_2^N(s_2 = 1, s_1 = 1 \| a_{j,y})$	1.00	0.80	0.75	0.70	0.75	0.65	0.75	1.00	1.00	1.00	1.00	1.00	0.85	0.80	0.85
$p_2^M(s_2 = 1, s_1 = 1 \| a_{j,y})$	0.00	0.17	0.20	0.25	0.20	0.30	0.22	0.00	0.00	0.00	0.00	0.00	0.15	0.20	0.15
$p_2^S(s_2 = 1, s_1 = 1 \| a_{j,y})$	0.00	0.03	0.05	0.05	0.05	0.05	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$p_2^N(s_1 = 0 \| a_{j,y})$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
$p_2^M(s_1 = 0 \| a_{j,y})$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$p_2^S(s_1 = 0 \| a_{j,y})$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

(III) Radiological effects

Probabilities	Attack strategies														
	$a_{1,1}$	$a_{1,2}$	$a_{1,3}$	$a_{1,4}$	$a_{2,1}$	$a_{2,2}$	$a_{2,3}$	$a_{3,1}$	$a_{3,2}$	$a_{3,3}$	$a_{3,4}$	$a_{4,1}$	$a_{4,2}$	$a_{4,3}$	$a_{4,4}$
$p_3^N(s_2 = 0, s_1 = 1 \| a_{j,y})$	1.00	0.75	0.60	0.80	0.50	0.50	0.50	1.00	1.00	1.00	1.00	1.00	0.95	0.60	0.95
$p_3^M(s_2 = 0, s_1 = 1 \| a_{j,y})$	0.00	0.20	0.30	0.15	0.40	0.40	0.40	0.00	0.00	0.00	0.00	0.00	0.05	0.30	0.05
$p_3^S(s_2 = 0, s_1 = 1 \| a_{j,y})$	0.00	0.05	0.10	0.05	0.10	0.10	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.00
$p_3^N(s_2 = 1, s_1 = 1 \| a_{j,y})$	1.00	0.95	0.95	0.95	0.95	0.95	0.95	1.00	1.00	1.00	1.00	1.00	0.95	0.95	0.95
$p_3^M(s_2 = 1, s_1 = 1 \| a_{j,y})$	0.00	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.00	0.00	0.00	0.00	0.05	0.05	0.05
$p_3^S(s_2 = 1, s_1 = 1 \| a_{j,y})$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

$p_3^N(s_1 = 0\|a_{j,y})$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
$p_3^M(s_1 = 0\|a_{j,y})$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$p_3^S(s_1 = 0\|a_{j,y})$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

(IV) Public panic and chaos

Probabilities	Attack strategies														
	$a_{1,1}$	$a_{1,2}$	$a_{1,3}$	$a_{1,4}$	$a_{2,1}$	$a_{2,2}$	$a_{2,3}$	$a_{3,1}$	$a_{3,2}$	$a_{3,3}$	$a_{3,4}$	$a_{4,1}$	$a_{4,2}$	$a_{4,3}$	$a_{4,4}$
$p_4^N(s_2 = 0, s_1 = 1\|a_{j,y})$	1.00	0.80	0.70	0.80	0.50	0.40	0.40	1.00	1.00	1.00	1.00	1.00	0.95	0.60	0.60
$p_4^M(s_2 = 0, s_1 = 1\|a_{j,y})$	0.00	0.15	0.25	0.17	0.40	0.50	0.40	0.00	0.00	0.00	0.00	0.00	0.05	0.20	0.20
$p_4^S(s_2 = 0, s_1 = 1\|a_{j,y})$	0.00	0.05	0.05	0.03	0.10	0.10	0.20	0.00	0.00	0.00	0.00	0.00	0.00	0.20	0.20
$p_4^N(s_2 = 1, s_1 = 1\|a_{j,y})$	1.00	0.90	0.85	0.90	0.70	0.75	0.70	1.00	1.00	1.00	1.00	1.00	0.95	0.90	0.90
$p_4^M(s_2 = 1, s_1 = 1\|a_{j,y})$	0.00	0.10	0.10	0.07	0.25	0.20	0.25	0.00	0.00	0.00	0.00	0.00	0.05	0.10	0.10
$p_4^S(s_2 = 1, s_1 = 1\|a_{j,y})$	0.00	0.00	0.05	0.03	0.05	0.05	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$p_4^N(s_1 = 0\|a_{j,y})$	1.00	0.98	0.98	0.98	0.95	0.95	0.95	1.00	1.00	1.00	1.00	1.00	0.98	0.95	0.95
$p_4^M(s_1 = 0\|a_{j,y})$	0.00	0.02	0.02	0.02	0.04	0.04	0.04	0.00	0.00	0.00	0.00	0.00	0.02	0.04	0.03
$p_4^S(s_1 = 0\|a_{j,y})$	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02

(V) Media impact

Probabilities	Attack strategies														
	$a_{1,1}$	$a_{1,2}$	$a_{1,3}$	$a_{1,4}$	$a_{2,1}$	$a_{2,2}$	$a_{2,3}$	$a_{3,1}$	$a_{3,2}$	$a_{3,3}$	$a_{3,4}$	$a_{4,1}$	$a_{4,2}$	$a_{4,3}$	$a_{4,4}$
$p_5^N(s_2 = 0, s_1 = 1\|a_{j,y})$	0.99	0.30	0.20	0.30	0.10	0.10	0.10	0.98	0.98	0.98	0.98	0.90	0.90	0.60	0.60
$p_5^M(s_2 = 0, s_1 = 1\|a_{j,y})$	0.01	0.65	0.70	0.65	0.70	0.70	0.70	0.02	0.02	0.02	0.02	0.10	0.10	0.30	0.30
$p_5^S(s_2 = 0, s_1 = 1\|a_{j,y})$	0.00	0.05	0.10	0.05	0.20	0.20	0.20	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.10
$p_5^N(s_2 = 1, s_1 = 1\|a_{j,y})$	0.99	0.85	0.60	0.60	0.50	0.50	0.60	0.98	0.98	0.98	0.98	0.95	0.95	0.80	0.70
$p_5^M(s_2 = 1, s_1 = 1\|a_{j,y})$	0.01	0.12	0.30	0.25	0.30	0.35	0.30	0.02	0.02	0.02	0.02	0.05	0.05	0.15	0.25
$p_5^S(s_2 = 1, s_1 = 1\|a_{j,y})$	0.00	0.03	0.10	0.05	0.20	0.15	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.05
$p_5^N(s_1 = 0\|a_{j,y})$	1.00	0.99	0.99	0.99	0.95	0.95	0.95	1.00	1.00	1.00	1.00	1.00	0.98	0.95	0.95
$p_5^M(s_1 = 0\|a_{j,y})$	0.00	0.01	0.01	0.01	0.04	0.04	0.04	0.00	0.00	0.00	0.00	0.00	0.02	0.04	0.03
$p_5^S(s_1 = 0\|a_{j,y})$	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02

4.1.2.2. The Adversarial Risk Analysis model

In particular, a defender is given advice on the optimal defense portfolio against cyber attacks, when only acquainted with subjective (partial) knowledge on attacker decisions.

Considering the outcomes of the game $\vec{s} = \{s_1(d_1^r, a_{j,y}), s_2(d_2^r|s_1)\}$ in the decision making, the defender seeks for the optimal resource allocation $d^* = \{d_1^*, d_2^*\}$ that is expected to optimally prevent the digital I&C system from unknown cyber attacks and, at the same time, minimize the system functionality loss in case of successful cyber attack. The d^* is obtained by maximizing the defender expected utility $\Psi_D(d^r)$:

$$d^* = \underset{d^r \in \Re}{\arg\max}\, \psi_D\left(d^r\right) \tag{4-20}$$

where $\Psi_D(d^r)$ is defined as in Eq. (4-21):

$$\psi_D\left(d^r\right) = \sum_{a_{j,y} \in A}\left[\sum_{s_2 \in \{0,1\}}\sum_{s_1 \in \{0,1\}} \pi_D\left(a_{j,y} \mid d^r\right) \cdot p_D\left(s_2 \mid d_2^r, s_1\right) \cdot p_D\left(s_1 \mid d_1^r, a_{j,y}\right) \cdot u_D\left(\vec{s} \mid d^r, a_{j,y}\right)\right] \tag{4-21}$$

where $u_D(\vec{s}|d^r, a_{j,y})$ is the defender utility of possible consequences costs, $\vec{p} = \left\{p\left(s_1(d_1^r, a_{j,y})\right), p(s_2(d_2^r|s_1))\right\}$ defines the defender assumptions on the probabilities of the outcomes (i.e., $s_1(d_1^r, a_{j,y})$ and $s_2(d_2^r|s_1)$), and $\pi_D(a_{j,y}|d^r)$ is the defender estimation of the probability of occurrence of any $a_{j,k}$ attack, given that the defense resources d^r are deployed.

To cope with the uncertainty on the type of attack (unknown to the defender) the MC approach sketched in Figure 28 is used for (a) estimating the $\pi_D(a_{j,y}|d^r)$ that is fundamental for (b) estimating the defender optimal defense strategy d^*.

(a) Estimation of $\boldsymbol{\pi_D(a_{j,y}|d^r)}$

The shadowed loop of Figure 28 (left) allows to mimic N_m different attacker decisions, and propagate the defender uncertainty on these decisions, with respect to a specific deployable d^r (from $d^1 = \{0,0,1,0,0,0,0\}$ to $d^{4834} = \{1,4,4,3,2,2,2\}$). At the m-th run, $m = 1, 2, \ldots, N_m$:

(a1) For each combination $(d_1^r, d_2^r, a_{j,y})$ given a d^r, sample the values of c_{prep}, c_{Aq} and k_A from the defender subjective distributions in Section 4.1.2.1, to calculate the attacker consequences of costs $c_A(\vec{s}|d^r, a_{j,y})$ of Eq. (4-17) and the corresponding utilities $u_A(\vec{s}|d^r, a_{j,y})$ of Eq. (4-19);

(a2) After sampling $p_A\left(s_1(d_1^r, a_{j,y})\right)$ and $p_A(s_2(d_2^r|s_1))$ from the defender subjective distributions in Section 4.1.2.1, calculate the attacker expected utility of $a_{j,k}$ conditioned on the d^r, $\Psi_A^m(a_{j,y}|d^r)$, by:

$$\psi_A^m\left(a_{j,y} \mid d^r\right) = \sum_{s_2\in\{0,1\}} \sum_{s_1\in\{0,1\}} p_A\left(s_2 \mid d_2^r, s_1\right) \cdot p_A\left(s_1 \mid d_1^r, a_{j,y}\right) \cdot u_A\left(\vec{s} \mid d^r, a_{j,y}\right) \tag{4-22}$$

(a3) Find the optimal attack strategy $a^{*,m}(d^r)$, with respect to the d^r:

$$a^{*,m}\left(d^r\right) = \arg\max_{a_{j,y}\in A} \psi_A^m\left(a_{j,y} \mid d^r\right) \tag{4-23}$$

(a4) Run N_m = 1000 time steps (a1) to (a3), to calculate $\vartheta(a_{j,y}|d^r)$ the number of $a_{j,k}$ being the optimal attack strategy at all the N_m runs, given the d^r;

(a5) Estimate $\pi_D(a_{j,y}|d^r)$ by:

$$\pi_D\left(a_{j,y} \mid d^r\right) = \frac{\vartheta\left(a_{j,y} \mid d^r\right)}{N_m} \tag{4-24}$$

(b) Estimation of d^* by a MC simulation

At the v-th run, v = 1, 2, …, N_v, of the MC simulation of Figure 28 (right),

(b1) For each one of the set of $\mathfrak{R}$ (=4834) defense portfolios, d^r, take the values of $\pi_D(a_{j,y}|d^r)$ (see (a)), with respect to each type of attacks $a_{j,y}$.

(b2) For each combination $(d_1^r, d_2^r, a_{j,y})$ given the d^r, sample the values of c_{annual}^r, $c_{Dl}(\vec{s}|d^r, a_{j,y})$, and k_D from the defender subjective distributions, respectively; taking the values of $\pi_D(a_{j,y}|d^r)$ of (a) and $\vec{p}_D = \left\{p_D\left(s_1(d_1^r, a_{j,y})\right), p_D(s_2(d_2^r|s_1))\right\}$ of Eqs. (4-9) and (4-10), calculate the defender expected utility $\Psi_D^v(d^r)$ by Eq. (4-21);

(b3) After calculating $\Psi_D^v(d^r)$ for all the portfolios d^r at the v-th run, find the optimal one by Eq. (4-20) that is equivalent to:

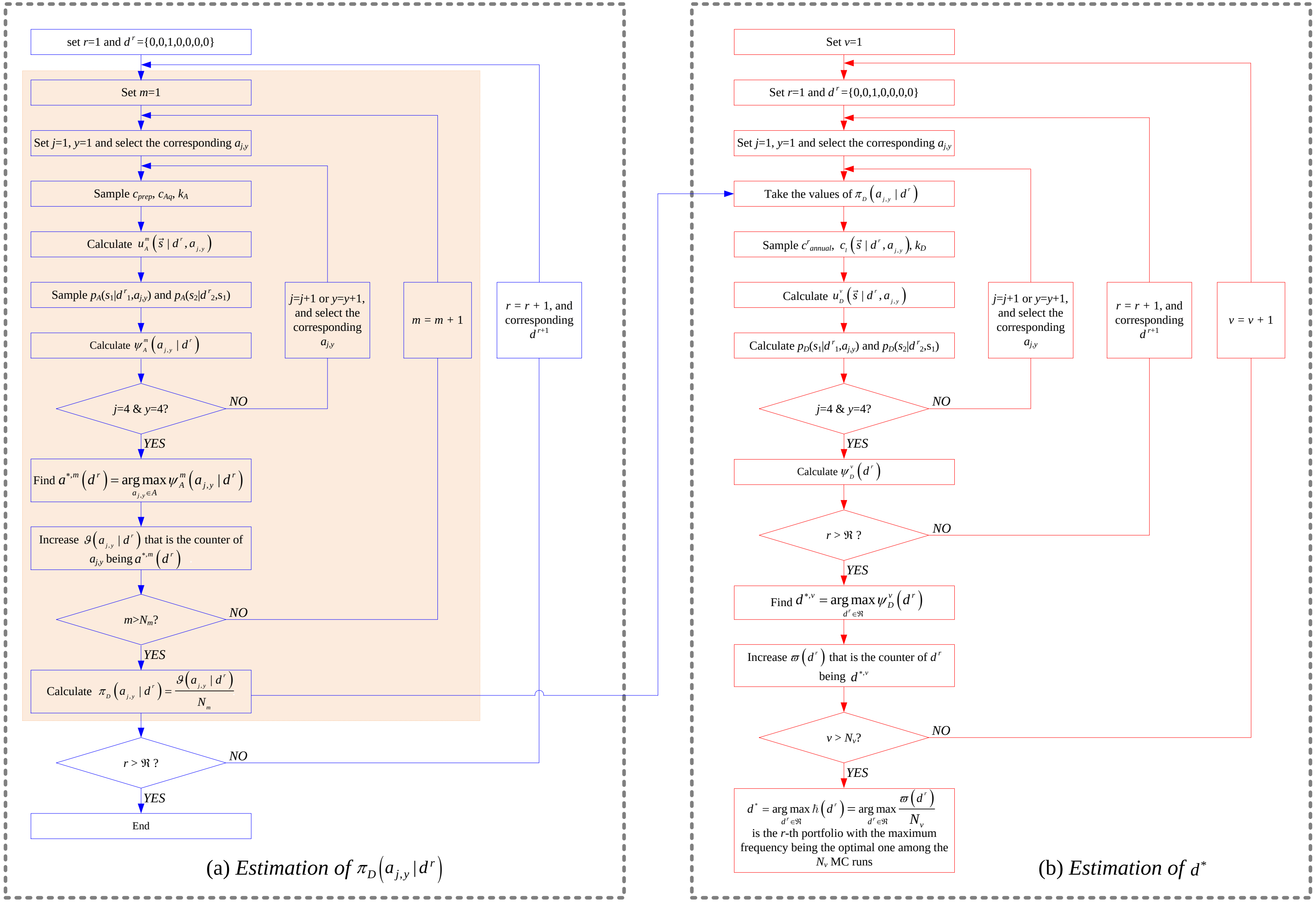
$$d^{*,v} = \arg\max_{d^r\in\mathfrak{R}} \psi_D^v\left(d^r\right) \tag{4-25}$$

(b4) Run N_v = 1000 times steps (b1) to (b3) to build the empirical $\hbar_D(d^r)$, which is the frequency of d^r being the optimal portfolio in all N_v runs;

(b5) Obtain the defender optimal resource allocation d^* that is:

$$d^* = \underset{d^r \in \Re}{\arg\max}\, \hbar_D\left(d^r\right) = \underset{d^r \in \Re}{\arg\max} \frac{\varpi\left(d^r\right)}{N_v} \tag{4-26}$$

where, $\varpi(d^r)$ is the number of times the d^r is the optimal portfolio in all N_v runs.


set r=1 and d^r={0,0,1,0,0,0,0}
Set m=1
Set j=1, y=1 and select the corresponding $a_{j,y}$
Sample c_{prep}, c_{Aq}, k_A
Calculate $u_A^m(\vec{s} \mid d^r, a_{j,y})$
Sample $p_A(s_1|d^r{}_1,a_{j,y})$ and $p_A(s_2|d^r{}_2,s_1)$
Calculate $\psi_A^m(a_{j,y} \mid d^r)$
j=4 & y=4?
NO
YES
j=j+1 or y=y+1, and select the corresponding $a_{j,y}$
$m = m + 1$
$r = r + 1$, and corresponding d^{r+1}
Find $a^{*,m}(d^r) = \arg\max_{a_{j,y}\in A} \psi_A^m(a_{j,y} \mid d^r)$
Increase $\vartheta(a_{j,y} \mid d^r)$ that is the counter of $a_{j,y}$ being $a^{*,m}(d^r)$
m>N_m?
NO
YES
Calculate $\pi_D(a_{j,y} \mid d^r) = \dfrac{\vartheta(a_{j,y} \mid d^r)}{N_m}$
$r > \Re$?
NO
YES
End
Set v=1
Set r=1 and d^r={0,0,1,0,0,0,0}
Set j=1, y=1 and select the corresponding $a_{j,y}$
Take the values of $\pi_D(a_{j,y} \mid d^r)$
Sample $c^r{}_{annual}$, $c_l(\vec{s} \mid d^r, a_{j,y})$, k_D
Calculate $u_D^v(\vec{s} \mid d^r, a_{j,y})$
Calculate $p_D(s_1|d^r{}_1,a_{j,y})$ and $p_D(s_2|d^r{}_2,s_1)$
j=4 & y=4?
NO
YES
j=j+1 or y=y+1, and select the corresponding $a_{j,y}$
$r = r + 1$, and corresponding d^{r+1}
$v = v + 1$
Calculate $\psi_D^v(d^r)$
$r > \Re$?
NO
YES
Find $d^{*,v} = \arg\max_{d^r\in\Re} \psi_D^v(d^r)$
Increase $\varpi(d^r)$ that is the counter of d^r being $d^{*,v}$
$v > N_v$?
NO
YES
$d^* = \arg\max_{d^r\in\Re} \hbar(d^r) = \arg\max_{d^r\in\Re} \dfrac{\varpi(d^r)}{N_v}$ is the r-th portfolio with the maximum frequency being the optimal one among the N_v MC runs

(a) *Estimation of* $\pi_D(a_{j,y} \mid d^r)$

(b) *Estimation of* d^*

Figure 28 The flowchart of the ARA approach for obtaining the optimal defense allocation

4.1.2.3. The classical defend-attack model

In most applications of traditional game theory, the attacker and the defender are assumed to share common knowledge regarding utility functions and probabilities of outcomes. Such assumption allows combining defender and attacker decision analysis into a coupled (balanced) model.

In game theory, the Nash equilibrium is defined to solve the players equilibrium strategies of a non-cooperative game involving two or more players and, can be reached if no player can do better by unilaterally changing his/her strategy [220, 234]. Each strategy in a Nash equilibrium is a best response providing a player with the most favorable outcome, taking other strategies in that equilibrium as given [235]. It has been proven that at least one Nash equilibrium exists for any game involving a finite number of players who can choose from finite strategies [236].

As proposed in [221], taking either player strategies and beliefs as given in the coupled defend-attack model, decision analysis allows reaching one player best response with respect to each of the opponent strategies and seeking an intersection point, namely, a Nash equilibrium, (d^*_{Nash}, a^*_{Nash}), that satisfies:

$$\psi_D\left(d^*_{Nash}, a^*_{Nash}\right) = \max_{d^r \in \Re} \psi_D\left(d^r, a^*_{Nash}\right) \quad \& \quad \psi_A\left(d^*_{Nash}, a^*_{Nash}\right) = \max_{a_{j,y} \in A} \psi_A\left(d^*_{Nash}, a_{j,y}\right) \qquad (4\text{-}27)$$

where, $a^*_{Nash}(d^r)$ is the attacker best response with respect to a defender decision d^r and $d^*_{Nash}(a_{j,y})$ is the defender best response with respect to an attacker strategy $a_{j,y}$, and they are obtained by Eqs. (4-28) and (4-29), respectively:

$$a^*_{Nash}\left(d^r\right) = \max_{a_{j,y} \in A} \psi_A\left(a_{j,y} \mid d^r\right) \qquad (4\text{-}28)$$

$$d^*_{Nash}\left(a_{j,y}\right) = \max_{d^r \in \Re} \psi_D\left(d^r \mid a_{j,y}\right) \qquad (4\text{-}29)$$

The Nash equilibrium (d^*_{Nash}, a^*_{Nash}) is commonly obtained as a combinatorial solution at which the defender and the attacker find a balanced strategy with the other, whereas, neither the defender nor the attacker can benefit from changing strategy with the other keeping the strategy unchanged [106, 220, 221].

4.1.3. Results

4.1.3.1. Optimal defensive resource allocation by Adversarial Risk Analysis

In ARA assessment, defender beliefs on the launching of an attack $a_{j,y}$ given a defense portfolio d^r, $\pi_D(a_{j,y}|d^r)$, can be estimated by MC simulation. On this basis, the defender optimal defense portfolio d^* is assesses by MC simulation, for taking into account the defender uncertainty on his/her predictive judgment on the countermeasure annual costs, the monetized consequences after attacks and the probabilities of outcomes.

As illustrative example, Figure 29 shows one run of the N_v estimates of the defender expected utilities of d^r (dots): the optimal defense portfolio is estimated as $d^{*,\nu} = \{1,3,4,2,2,2,2\}$ (diamond) with the (absolute) lowest value of expected utilities $\Psi_D(d^{*,\nu})$ equal to 1.0753 (i.e., the defender reaches the setting of lowest expected investment against the uncertain attacks, with an attitude of risk aversion) and the countermeasures annual costs equal to 1,865 k€, under an assumed B_M equal to 2,000 k€ (for sake of illustration).

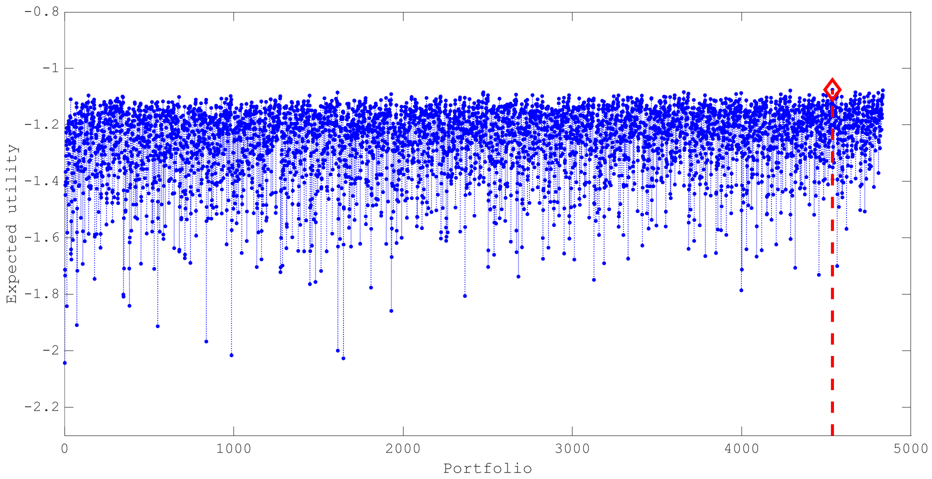


Figure 29 The defender's expected utilities with respect to each portfolio

It can be seen that many other portfolios reach expected utilities close to $\Psi_D(d^*)$. In Table 28, the top five defense portfolios with highest utility values are listed. As a matter of fact, even though the utility estimates are similar, countermeasures portfolios change much, supporting the need of a robust approach to provide the optimal result with the needed confidence.

Table 28 The optimal defense portfolios with annual costs

d^r	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$x_{1,4}$	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$\Psi_D(d^r)$	c^r_{annual} (k€)
	$n^r_{1,1}$	$n^r_{1,2}$	$n^r_{1,3}$	$n^r_{1,4}$	$n^r_{2,1}$	$n^r_{2,2}$	$n^r_{2,3}$		
d^{4345}	1	3	4	2	2	2	2	-1.0753	1,865
d^{4834}	1	4	4	3	2	2	2	-1.0773	1,960
d^{4779}	1	4	4	1	2	2	2	-1.0780	1,800
d^{4260}	1	3	3	3	2	2	2	-1.0846	1,895
d^{1997}	0	3	4	3	2	2	2	-1.0843	1,865

In line with the proposed approach, therefore, the N_v runs of MC simulation lead to $d^* = d^{4779} = \{1,4,4,1,2,2,2\}$ (diamond in Figure 30) with the largest value of the frequency of d^r being the optimal portfolio $\hbar_D(d^r)$ equal to 7e-3 for the optimal defense portfolio, that is taken as confidence measure for the result provided, leveraging the robustness of the protection actions on the ALFRED digital I&C system with uncertain malicious threats characteristics.

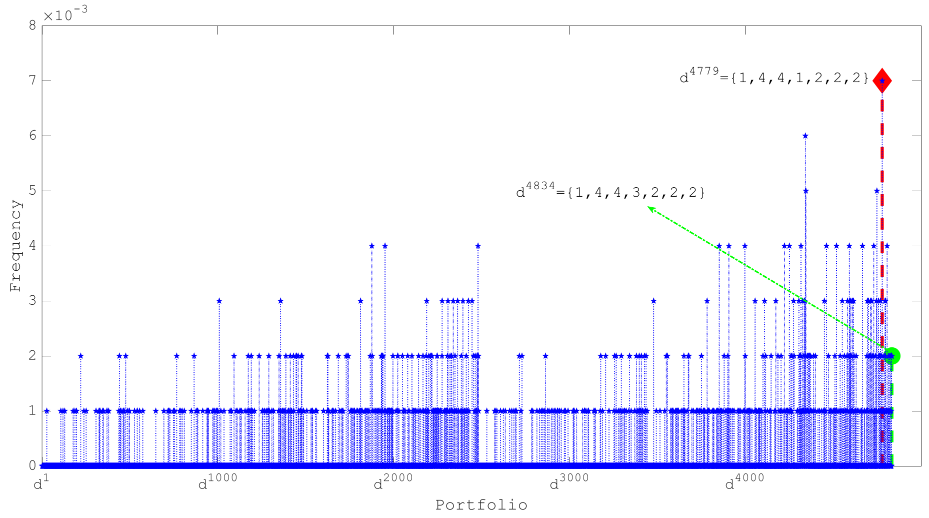


Figure 30 Optimal defense portfolio from ARA assessment

It is worth mentioning that $d^{4342} = \{1,3,4,2,2,1,2\}$, $d^{4749} = \{1,4,4,0,2,2,2\}$ and $d^{4345} = \{1,3,4,2,2,2,2\}$ turn out to be sub-optimal portfolios since they are estimated as $d^{*,v}$ among the N_v runs for 6, 5 and 5 times, as listed in Table 29, respectively. This suggests that the development and maintenance of a security software is usually time-consuming but may impair the CPS security level (if properly designed) (for example, the security analyst would be more likely to select d^{4779} (equipped with $n^r_{1,4} = 1$ security software ($x_{1,4}$)) but not d^{4749} (without security software (i.e., $n^r_{1,4} = 0$) for defense resource allocation), whereas, operators devoted to real-time monitoring of physical processing are more likely prone to human errors (for example, it is impossible to recruit only one operator in NPPs, as shown in d^{3998} (i.e., $n_{1,3}^{r}$ = 1)).

Table 29 The optimal defense portfolios with annual costs

d^r	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$x_{1,4}$	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$\Psi_D(d^r)$	c_{annual}^{r}
	$n_{1,1}^{r}$	$n_{1,2}^{r}$	$n_{1,3}^{r}$	$n_{1,4}^{r}$	$n_{2,1}^{r}$	$n_{2,2}^{r}$	$n_{2,3}^{r}$		(k€)
d^{4779}	1	4	4	1	2	2	2	7e-3	1,800
d^{4342}	1	3	4	2	2	1	2	6e-3	1,795
d^{4749}	1	4	4	0	2	2	2	5e-3	1,720
d^{4345}	1	3	4	2	2	2	2	5e-3	1,865
d^{3998}	1	3	1	2	2	2	2	4e-3	1,715

4.1.3.2. The Nash equilibrium solution from the classical defend-attack model

In the model of Section 4.1.2.3, the attacker beliefs, i.e., probabilities of outcomes, costs, risk aversion coefficient, are known to the defender, and assumed to be mean values from the corresponding distributions mentioned in Section 4.1.2.1.

Figure 31 shows the Nash equilibrium solution $(d_{Nash}^{*}, a_{Nash}^{*})$ obtained by finding the intersection node between the defender best responses with respect to $a_{j,y}$ and the attacker best responses with respect to d^r. On one hand, the attacker calculates his/her expected utility of each attack decision $a_{j,y}$ (of the pool of A) $\pi_D(a_{j,y}|d^r)$ and obtains a best response $a_{Nash}^{*}(d^r)$ with respect to a defender strategy d^r (out of $\mathfrak{R}$ = 4834 portfolios). The solutions of Eq. (4-28) turn out to be the constant attacker best response $a_{2,1}$ (attack to control rod actuator) with respect to different defense strategies (see dot line in Figure 31). Whereas, on the other hand, the defender calculates his/her expected utility of each d^r (out of $\mathfrak{R}$ = 4834 portfolios) $\pi_D(a_{j,y}|d^r)$ and obtains a best response $d_{Nash}^{*}(a_{j,y})$ with respect to an attack decision $a_{j,y}$ (of A). The solutions of Eq. (4-29) vary with the attack decisions (see stars in Figure 31). Notably, the attacker best responses $a_{Nash}^{*}(d^r)$ and the defender best responses $d_{Nash}^{*}(a_{j,y})$ intersect at the point of $(d^{4834}, a_{2,1})$, and Nash equilibrium is $(d_{Nash}^{*}, a_{Nash}^{*})$ (see diamond in Figure 31), where $d^{4834} = \{1,4,4,3,2,2,2\}$ is equipped with all deployable defensive resources under the restriction of B_M equal to 2,000 k€.

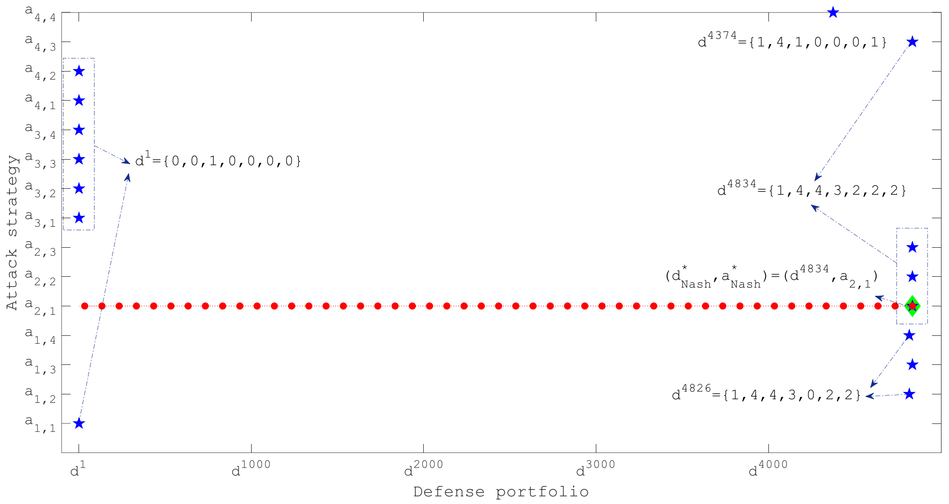


Figure 31 Estimation of Nash equilibrium solution from the classical defend-attack model

It must be noticed that the Nash equilibrium solution $d_{Nash}^{*} = d^{4834} = \{1,4,4,3,2,2,2\}$ (shown with a circle in Figure 30), obtained from the classical defend-attack model that assumes that the defender and the attacker share common knowledge, differs from $d^{*} = d^{4779} = \{1,4,4,1,2,2,2\}$ by two sets of security softwares (i.e., $n_{1,4}^{4834} - n_{1,4}^{4779} = 2$).

Even if surprisingly marginal, there is indeed a fundamental difference between the two solutions: the Nash equilibrium solution d^{4834} in practice assumes the maximum quantity of defense resources to be installed with the maximum allowed budget B_M, whereas, the optimal decision of the ARA d^{*} (i.e., d^{4779} highlighted in diamond in Figure 30) reaches the one-sided prescriptive optimal decision against all possible uncertain cyber attacks without reaching the maximum budget. Moreover, as shown in Figure 30, the allocation strategy d^{4834} gives a value of $\hbar_D(d^{*})$ equal to 2e-3 and, therefore, less effective in protecting the CPS from the uncertain attacks than d^{4779}.

4.1.4. Conclusions and discussions

In this Chapter, we have proposed an ARA approach for analyzing decisions between intelligent adversaries provided a novel one-sided (i.e., defender) prescriptive support strategy for optimizing the defensive resource allocations based on a subjective expected utility model.

A MC approach has been embedded into the ARA model for treating uncertainties in the decisions of the adversaries, for improving confidence in obtaining the optimal defense resource allocation, leveraging robustness of protection actions on the CPS with uncertain malicious threats.

For demonstration, we have illustrated the proposed ARA framework to a cyber defend-attack game in the digital I&C system of the ALFRED (Paper 4[J]). With respect to the prescriptive support, the ARA framework advised the defender the optimal portfolio of defense resource allocation, minimizing the system integrity loss against uncertain cyber attacks. The result has also been compared with the Nash equilibrium solution from a classical defend-attack model, in which the attacker and the defender share common knowledge regarding utility functions and probabilities of the outcomes of the game, showing that a stable status (Nash equilibrium) can be reached between the defender and the attacker, as a two-sided prescriptively balanced strategy profile.

It is worth noting that, in reality, the security defender seems to be more likely to protect and recover a CPS relying on the cyber security emergency management from a launched specific cyber attack [237], rather than solely on the prevention of the optimized allocation of resources for defensive barriers referred from a probabilistic analysis (e.g., the proposed ARA modelling). Despite this, decision analysis on the defensive resource allocation is still of importance, specifically to highly (cyber) defended facilities such as NPPs, for exploring and understanding how the cyber attacks impact the system functionality and its environment, that is fundamental to plan proper protection and mitigation actions for resilience, especially under a constrained defense budget.

4.2. A Non-Parametric Cumulative Sum Approach for online diagnostic of cyber attacks

Both stochastic failures and cyber attacks can compromise the correct functionality of Cyber-Physical Systems (CPSs). Cyber attacks manifest themselves in the physical system and, can be misclassified as component failures, leading to wrong control actions and maintenance strategies. In this Chapter (i.e., Task IV.2), we illustrate the use of the NP-CUSUM approach for online diagnostics of cyber attacks to CPSs. This allows for promptly recognizing cyber attacks by distinguishing them from component failures, and guiding decisions for the CPSs recovery from anomalous conditions.

4.2.1. Case study: the ALFRED

We apply the approach to the ALFRED and its digital I&C system. Redundancy is commonly applied to sensors and signal processing units of a digital I&C system [238]. In the ALFRED digital control scheme, redundancy has been used to design each independent SISO loop. Figure 32 shows an example of the redundant design scheme of the $T_{L,cold}$-PI$_3$-G_{water} control loop. The real values of the coolant SG outlet temperature $T_{L,cold}(t)$ are measured by a sensor. After collected and converted to quantized (discretized) values by a data acquisition system, the measurements are duplicated by two identical digital-to-analog converters (DACs) to Subsystem 1 for computing (feeding) and 2 for monitoring, respectively. The received measurements of Subsystem 1 $T_{L,cold}^{feed}(t)$ is fed to the computational unit PI$_3$, whereas those of Subsystem 2 $T_{L,cold}^{monitor}(t)$ are taken as redundant data, for detecting anomalous conditions of the physical system.

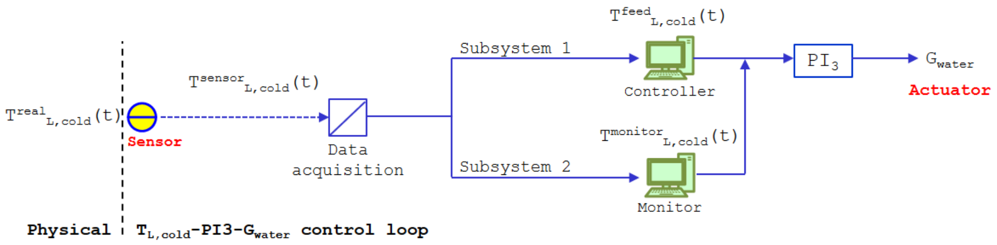


Figure 32 The redundancy design of the $T_{L,cold}$-PI3-G_{water} control loop

Measurements are realistically considered to be affected by two types of errors [239, 240]: measurement errors (assumed distributed according to a normal distribution) and quantization errors (which are rooted in the DACs and are assumed uniformly distributed between −1/2 and +1/2 Least Significant Bit (LSB)). For simplicity, but without loss of realism, Table 30 lists the reference values of the controlled variables, the distributions of sensor measurement errors and the quantization errors that each control loops is subjected to.

Table 30 List of reference parameters for safety variables

Variable, y	Reference value, y^{ref}, at full power nominal conditions	Sensor measuring error $\delta_y(t)$	Converters quantization error $q_y(t)$
T_{steam} (°C)	450	$N(0,1)$	[-0.05, +0.05]
p_{SG} (Pa)	$180 \cdot 10^5$	$N(0,0.1) \cdot 10^5$	$[-0.01,+0.01] \cdot 10^5$
$T_{L,cold}$ (°C)	400	$N(0,1)$	[-0.05, +0.05]
P_{Th} (W)	$300 \cdot 10^6$	$N(0,0.5) \cdot 10^6$	$[-0.05,+0.05] \cdot 10^6$

4.2.1.1. Failures

Without loss of generality, we hereby consider only the $T_{L,cold}$ sensor failures (but the following discussion remains valid for any other sensor of the I&C system of the ALFRED). The occurrence of sensor failure (i.e., bias, drift, wider noise and freezing [53]) at random time t_R results in altered sensor measurement $y^{sensor}(t)$ and, then, potentially lead the ALFRED to accidents [203, 241, 242], equivalent to the consequences of cyber attacks to sensors mentioned in Eq. (3-1).

Stochastic failures cause differences of the measurements $T_{L,cold}^{sensor}(t)$ from the real values of the controlled variable in the physical system. The MC sampling procedure used to inject stochastic failures to the $T_{L,cold}$ sensor at uniform random time t_R consists in sampling the sensor failures at random magnitudes in different failure modes and, then, running the ALFRED simulator for generating the controlled variables evolution throughout the mission time t_M. Erroneous measurements are, then, converted to two sets of quantized data in the data acquisition system and fed to both the computing (feeding) and monitoring subsystems, as shown in Figure 33.

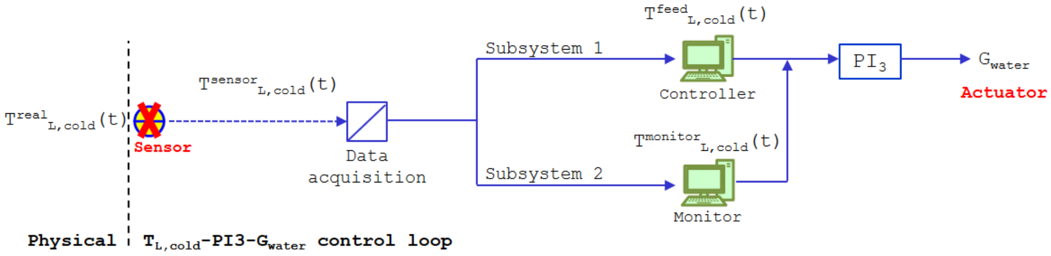


Figure 33 Schematics of $T_{L,cold}$ sensor stochastic failures

4.2.1.2. Cyber breaches

Alternatively, a DoS attack of Eq. (3-1) is modelled to block a legitimate packet traffic that processes the genuine connection and is substituted by a malicious packet traffic, preventing the controllers from receiving legitimate measurements and mimicking the stochastic sensor failures of Figure 33. Figure 34 shows the schematics of a DoS attack, in which the computing unit is fed by malicious packet traffic, altering the legitimate information, whereas, a legitimate packet traffic is regularly fed to the monitoring unit. DoS attacks are modelled to occur at uniform random time t_R within the time horizon t_M, and the sensor failures previously explained are mimicked.

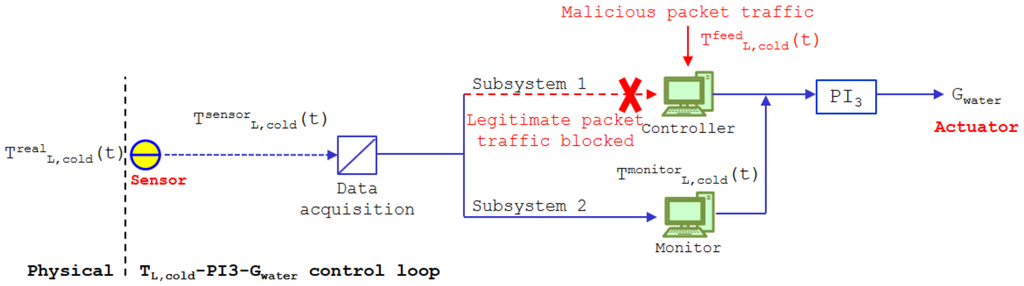


Figure 34 Schematics of DoS attacks

Both sensor failures and DoS attacks occur at unknown times, leading to unpredictable changes in the distributions of physical variables that differ from the normal condition distribution. Cyber attacks manifest themselves in the physical system and, can be misclassified as component failures, leading to wrong control actions and maintenance strategies. In this sense, diagnostic of cyber attacks and component failures is important for the system protection and resilience, allowing prompt recovery from the effects of disruptive events and, thus, increasing system resilience.

4.2.2. The NP-CUSUM online diagnostic tool

In this work, an online diagnostic tool based on a NP-CUSUM algorithm is embedded within the redundancy design of the control loops (e.g., illustrated in the $T_{L,cold}$-PI$_3$-G_{water} control loop, without loss of generality), for distinguishing between the stochastic sensor failures and the DoS attacks. As shown in Figure 35, the online diagnostic tool involves two main functions: (i) on-line collection of measurements received by the controllers, which are fed to the NP-CUSUM algorithm that is (off-line) trained on different system behaviors to set its parameters; (ii) an on-line application of the rules of classification of failures and cyber attacks.

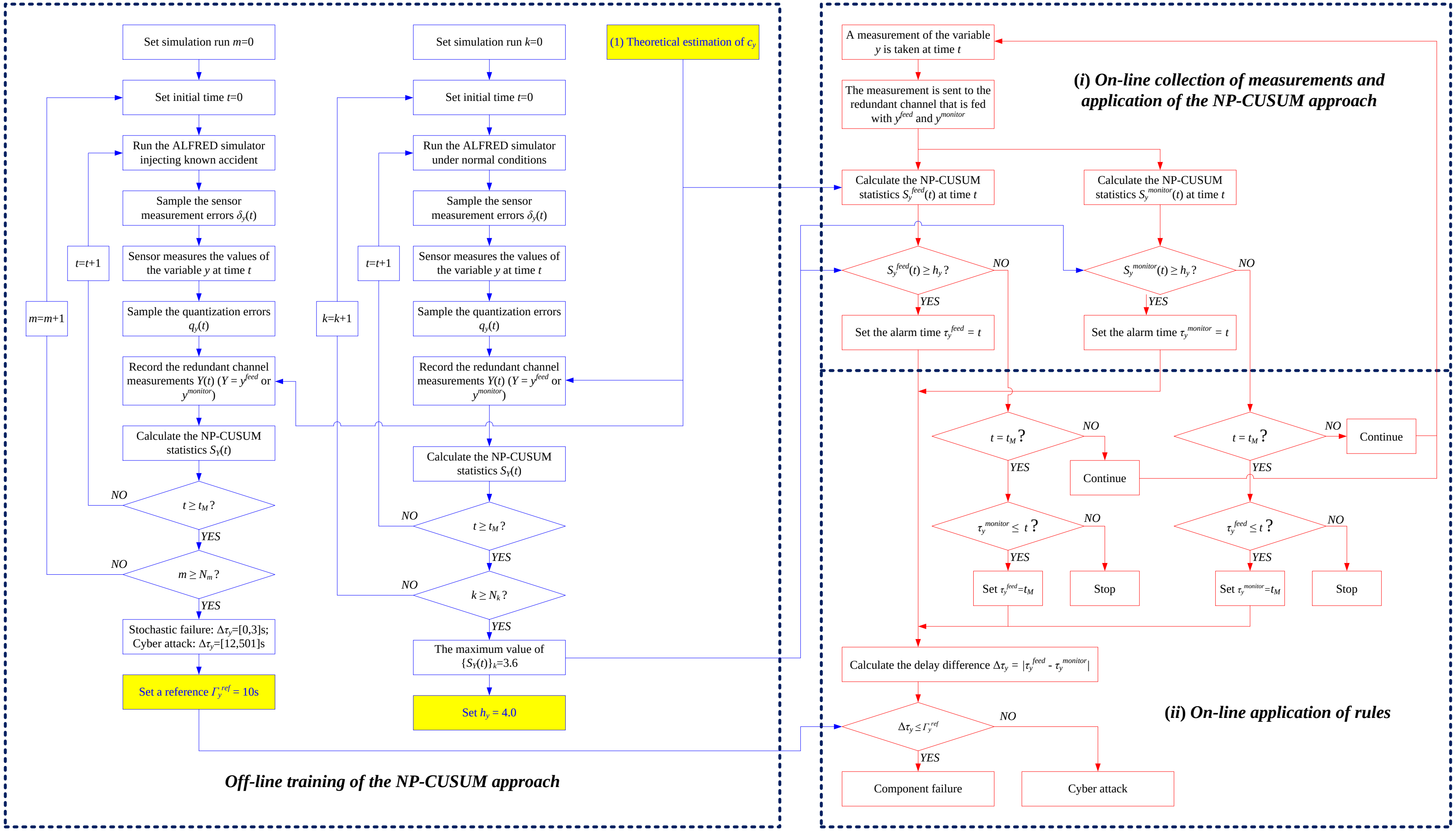

Set simulation run m=0
Set initial time t=0
Run the ALFRED simulator injecting known accident
Sample the sensor measurement errors δy(t)
Sensor measures the values of the variable y at time t
Sample the quantization errors qy(t)
Record the redundant channel measurements Y(t) (Y = yfeed or ymonitor)
Calculate the NP-CUSUM statistics SY(t)
t=t+1
m=m+1
NO
t ≥ tM ?
YES
NO
m ≥ Nm ?
YES
Stochastic failure: Δτy=[0,3]s; Cyber attack: Δτy=[12,501]s
Set a reference Γyref = 10s
Set simulation run k=0
Set initial time t=0
Run the ALFRED simulator under normal conditions
Sample the sensor measurement errors δy(t)
Sensor measures the values of the variable y at time t
Sample the quantization errors qy(t)
Record the redundant channel measurements Y(t) (Y = yfeed or ymonitor)
Calculate the NP-CUSUM statistics SY(t)
t=t+1
k=k+1
NO
t ≥ tM ?
YES
NO
k ≥ Nk ?
YES
The maximum value of {SY(t)}k=3.6
Set hy = 4.0
(1) Theoretical estimation of cy
Off-line training of the NP-CUSUM approach
A measurement of the variable y is taken at time t
The measurement is sent to the redundant channel that is fed with yfeed and ymonitor
(i) On-line collection of measurements and application of the NP-CUSUM approach
Calculate the NP-CUSUM statistics Syfeed(t) at time t
Calculate the NP-CUSUM statistics Symonitor(t) at time t
Syfeed(t) ≥ hy ?
NO
YES
Symonitor(t) ≥ hy ?
NO
YES
Set the alarm time τyfeed = t
Set the alarm time τymonitor = t
t = tM ?
NO
YES
Continue
t = tM ?
NO
YES
Continue
τymonitor ≤ t ?
NO
YES
Set τyfeed=tM
Stop
τyfeed ≤ t ?
NO
YES
Set τymonitor=tM
Stop
Calculate the delay difference Δτy = |τyfeed - τymonitor|
Δτy ≤ Γyref
NO
YES
Component failure
Cyber attack
(ii) On-line application of rules

Figure 35 Flowchart of the NP-CUSUM diagnostic approach

(*i*) On-line collection of measurements and application of the NP-CUSUM approach

The redundant channel measurements $Y(t)$, Y=y^{feed} and $y^{monitor}$, where $y = T_{L,cold}$, are collected online by the subsystems as follows. At each time t,

(1) The sensor measures the values $y = T_{L,cold}$, which is affected by the sensor measurement error $\delta_y(t)$ distributed as a normal distribution of Table 30, i.e., $y^{sensor}(t) = y^{real}(t) + \delta_y(t)$;

(2) The data acquisition system collects and converts $y^{sensor}(t)$ with the quantization accuracy $q_y(t)$ of Table 30, resulting in two redundant channels of quantized measurements;

(3) The computing and monitoring subsystems receive the redundant measurements $Y(t)$;

(4) The NP-CUSUM algorithm calculates score function-based statistics $S_Y(t)$ of the collected $Y(t)$, to check whether either $S_y^{feed}(t)$ or $S_y^{monitor}(t)$ exceeds a predefined threshold h_y:

- If yes, record the time to alarm τ_Y (τ_y^{feed} or/and $\tau_y^{monitor}$, respectively, and proceed with the rule-based diagnostics at Step (*ii*));
- If either $S_y^{feed}(t)$ or $S_y^{monitor}(t)$ exceeds h_y, collect the successive measurement because the monitored component is working under normal conditions.

(*ii*) On-line application of rules

(5) If both τ_y^{feed} and $\tau_y^{monitor}$ exist, calculate the delay difference $\Delta\tau_y$ (i.e., denoting the difference between the time-to-detection delays τ_y^{feed} and $\tau_y^{monitor}$):

$$\Delta\tau_y = \left|\tau_y^{feed} - \tau_y^{monitor}\right| \tag{4-30}$$

Otherwise, set $\tau_y^{monitor}$ equal to t_M (when $S_y^{monitor}(t)$ has not exceeded h_y when $S_y^{feed}(t)$ does, and vice versa, respectively).

If neither exists before t_M, continue diagnostics.

(6) Compare $\Delta\tau_y$ with a predefined reference delay difference Γ_y^{ref} and take decision:

- If $\Delta\tau_y \leq \Gamma_y^{ref}$, classify the event as ***Failure***;

- If $\Delta\tau_y > \Gamma_y^{ref}$, classify the event as ***Cyber Attack***.

The reference delay difference Γ_y^{ref} is estimated on a batch of N_m = 100 reference simulations, where, for each m-th simulation, a known component failure or cyber attack is injected. The minimum and maximum collected values of $\Delta\tau_y$ are found to be equal to 0s and 3s in case of components failures, and 12s and 501s in case of cyber attacks. Thus, we conservatively set Γ_y^{ref} equal to 10s, so that $\Delta\tau_y$ larger than 10s indicates that a cyber attack has occurred on the feeding subsystem.

(*iii*) Off-line training of the NP-CUSUM algorithm

The NP-CUSUM algorithm requires that the parameters c_y and h_y be customized to the different system behaviors, to guarantee the maximum capability of discriminating between failures and cyber attacks, in the ALFRED system.

(1) Estimation of c_y

A positive constant of c_y needs to be set in such a way to guarantee a negative mean value of $\mu_{\Delta g_Y} = \sum_t \Delta g_Y\left(Y(t)\right)\big/t$, $t = dt, 2dt, \ldots, t, (t<t_R)$, to hold before any anomaly (either failure or cyber attack) is detected, and a positive mean value $\theta_{\Delta g_Y} = \sum_t \Delta g_Y\left(Y(t)\right)\big/(t-t_R)$, $t = t_R, t_R+dt, t_R+2dt, \ldots$, to hold after the anomaly occurrence [122], viz:

$$\mu_{\Delta g_Y} = E\left[\omega_y \cdot \left(\left|Y(t)-\mu_Y\right| - c_y\right)\right] = -\omega_y \cdot \left(\frac{2\sigma_Y}{\sqrt{2\pi}} - c_y\right) < 0 \qquad (4\text{-}31)$$

$$\theta_{\Delta g_Y} \geq \omega_y \cdot \left(\left|\hat{\theta}_Y(t) - \mu_Y\right|_{\min} - c_y\right) > 0 \qquad (4\text{-}32)$$

where, $\left|\hat{\theta}_Y - \mu_Y\right|_{\min}$ is defined as the minimum difference between the estimated post-change mean $\hat{\theta}_{\Delta g_Y}$ and the known pre-change mean $\mu_{\Delta g_Y}$. As a result,

$$\frac{2\sigma_Y}{\sqrt{2\pi}} < c_y < \left|\hat{\theta}_Y(t) - \mu_Y\right|_{\min} \qquad (4\text{-}33)$$

where,

$$c_y = \varepsilon_y \cdot \hat{\theta}_{Y,a} \tag{4-34}$$

where $\hat{\theta}_{Y,a}$ is a postulated post-change mean value for an accidental scenario a.

Since under normal conditions, the probability of $Y(t)$ (distributed according to a normal distribution $N(\mu_Y, \sigma_Y)$) of falling within the interval $[\mu_Y\text{-}2\sigma_Y, \mu_Y+2\sigma_Y]$ is at least equal to 0.95 [243], viz:

$$\Pr\left[\mu_Y - 2\sigma_Y \le Y(t) \le \mu_Y + 2\sigma_Y\right] \ge 0.95 \tag{4-35}$$

we assume an anomaly to be occurred if $\hat{\theta}_{Y,a}$ falls outside the interval $[\mu_Y\text{-}2\sigma_Y, \mu_Y+2\sigma_Y]$. Without loss of generality, we suppose that $\hat{\theta}_{Y,a} > \mu_Y$. The minimum value of $\hat{\theta}_{Y,a}$ results to be equal to $\mu_Y+2\sigma_Y$ and, thus, $\left|\hat{\theta}_{Y,a} - \mu_Y\right|_{\min}$ is equal to $2\sigma_Y$. Eqs. (4-33) and (4-34) change to:

$$\frac{1}{\sqrt{2\pi}}\left(1 - \frac{\mu_Y}{\mu_Y + 2\sigma_Y}\right) < \varepsilon_y < 1 - \frac{\mu_Y}{\mu_Y + 2\sigma_Y} \tag{4-36}$$

In conclusion, without loss of generality, we take a value of ε_Y equal to:

$$\varepsilon_y = \frac{1}{2}\left(1 - \frac{\mu_Y}{\mu_Y + 2\sigma_Y}\right) \tag{4-37}$$

that, with respect to ($T_{L,cold}$ distributed as $N(400,1)$°C) makes c_y turn out to be equal to 1.005°C.

(2) Estimation of h_y

The threshold h_y can be set relying on a batch of N_k reference simulations under normal conditions, whose behaviors of the variable y without change points to failures or cyber attacks can be learnt, the NP-CUSUM statistics calculated and the parameter tailored to the simulation results. Specifically, we utilize N_k = 100 ALFRED randomly generated simulations. For each k-th simulation,

(a) Record the redundant channel measurements, $Y(t)$, $Y = y^{feed}$ or $y^{monitor}$, at each time t, $t = dt, 2dt, \ldots, t_M$;

(b) Calculate the corresponding NP-CUSUM statistics, $S_Y(t)$.

(c) Set the threshold h_y such that:

$$h_y > \max_{1 \le k \le N_k} \left\{ h_{y,k} \right\} \tag{4-38}$$

where,

$$h_{y,k} = \max_{1 \le t \le t_M} \left\{ S_Y \left(t \right) \right\}_k \tag{4-39}$$

and, $\{S_Y(t)\}_k$ is the collection of the statistics for the k-th simulation.

As shown in Figure 36 with respect to $T_{L,cold}$, the maximum value of the NP-CUSUM statistics is equal to 3.6 and, therefore, in what follows, we conservatively set $h_{T_{L,cold}}$ equal to 4.0.

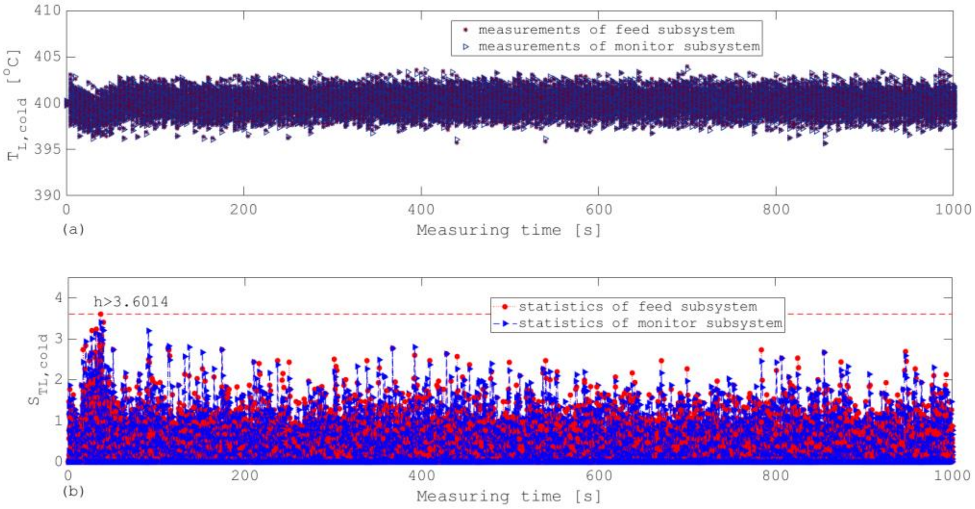


Figure 36 Estimation of the threshold $h_{T_{L,cold}}$: (a) the received measurements of the two subsystems of the control loop; (b) the corresponding statistics calculated from the measurements

4.2.3. Results

We illustrate the results of the NP-CUSUM-based diagnostic approach considering different $T_{L,cold}$ sensor failures and cyber attacks to the $T_{L,cold}$-PI_3-G_{water} control loop.

4.2.3.1. Diganosis of attacks mimicking bias failure mode

As illustration, Figure 37 presents the results of injecting bias failure at time t_R = 630s with a bias factor b equal to 7.569°C, leading the $T_{L,cold}$ sensor measurements $T_{L,cold}^{F,sensor}(t)$ to $T_{L,cold}(t) + b + \delta_{T_{L,cold}}(t)$, where $t \ge t_R$. As shown in Figure 37(a), the $T_{L,cold}$ sensor bias failure deviates both measurements $T_{L,cold}^{feed}(t)$ and $T_{L,cold}^{monitor}(t)$ from the real values of the physical system $T_{L,cold}(t)$. Figure 37 shows that the bias results in very quick response of both statistics evaluated on the measurements $T_{L,cold}^{feed}(t)$ and $T_{L,cold}^{monitor}(t)$: both statistics reach quickly the threshold $h_{T_{L,cold}}$ (dotted line) and the difference $\Delta\tau_{T_{L,cold}}$ between times to alarm ($\tau_{T_{L,cold}}^{feed}$ and $\tau_{T_{L,cold}}^{monitor}$) turns out to be equal to 0 (i.e., less than $\Gamma_{T_{L,cold}}^{ref}$ equal to 10s) (see Figure 37(b)), allowing for a (correct) identification of the event as a sensor failure mode and not as a cyber attack.

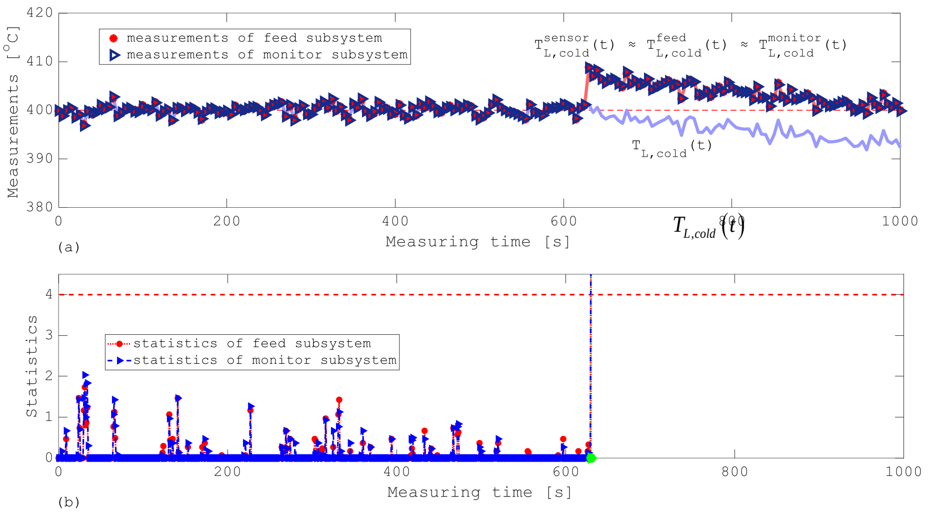


Figure 37 $T_{L,cold}$ sensor bias failure mode: (a) the received measurements of feed and monitor Subsystems in which the bias occurs at time t_R equal to 630s; (b) the corresponding NP-CUSUM statistics for diagnosing the bias failure

Contrarily, Figure 38(a) shows a cyber attack to the computing unit mimicking a bias failure mode at t_R=630s (with b again equal to 7.569°C): this leads $T_{L,cold}^{feed}(t)$ to deviate from $T_{L,cold}^{monitor}(t)$ (that, indeed, is the legitimate $T_{L,cold}^{sensor}(t)$ measured by the $T_{L,cold}$ sensor). The different values between the malicious and the legitimate measurements, then, lead to a delay response $\Delta\tau_{T_{L,cold}}$ equal to 66s (larger than $\Gamma_{T_{L,cold}}^{ref}$) between the threshold exceedance of $S_{T_{L,cold}}^{feed}(t)$ and $S_{T_{L,cold}}^{monitor}(t)$ (see Figure 38(b)), and allowing for a (correct) identification of the event as a cyber attack.

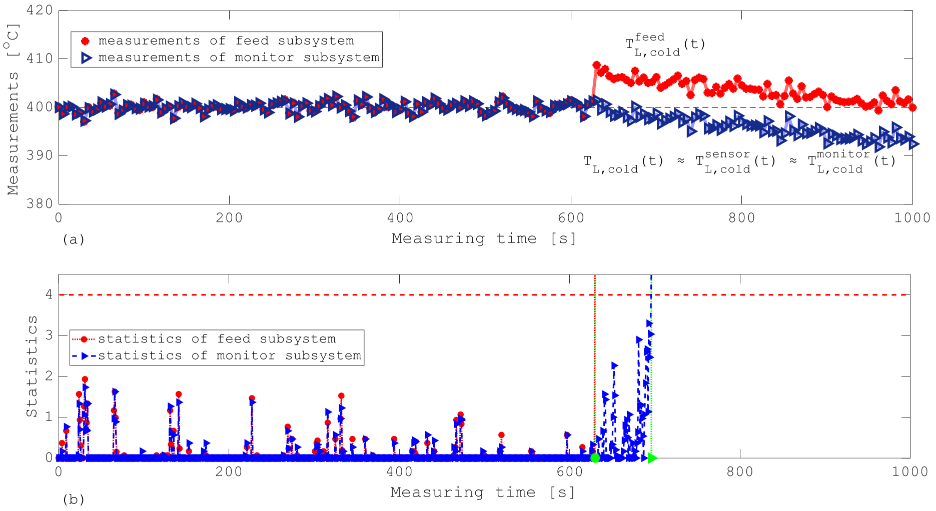


Figure 38 Cyber attack to the computing unit mimicking a bias failure mode: (a) the received measurements of feed and monitor Subsystems in which the cyber attack occurs at time t_R equal to 630s; (b) the corresponding NP-CUSUM statistics for diagnosing the cyber attack

4.2.3.2. *Diagnosis of attacks mimicking drift failure mode*

Figure 39 presents the results of injecting a drift at time t_R = 740s, with the drift factor c equal to 0.398. The drift c results in a very quick response of both statistics evaluated on the measurements $T^{feed}_{L,cold}(t)$ and $T^{monitor}_{L,cold}(t)$: both statistics reach quickly the threshold $h_{T_{L,cold}}$ (dotted line) and the difference $\Delta\tau_{T_{L,cold}}$ between times to alarm ($\tau^{feed}_{T_{L,cold}}$ and $\tau^{monitor}_{T_{L,cold}}$) turns out to be equal to 0 (i.e., less than $\Gamma^{ref}_{T_{L,cold}}$) (see Figure 39(b)), allowing for a (correct) identification of the event as a sensor failure.

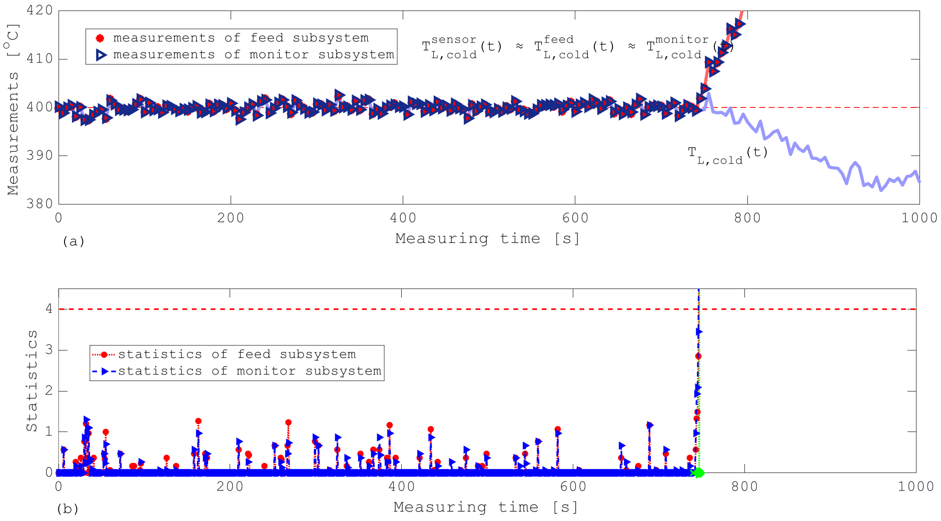


Figure 39 $T_{L,cold}$ sensor drift failure mode: (a) the received measurements of feed and monitor Subsystems in which the drift occurs at time t_R equal to 740s; (b) the corresponding NP-CUSUM statistics for diagnosing the bias failure

Contrarily, Figure 40(a) shows a cyber attack to the computing unit mimicking a drift failure mode at t_R=740s (with c again equal to 0.398), leading $T_{L,cold}^{feed}(t)$ to deviate from the legitimate $T_{L,cold}^{monitor}(t)$. The different values between the malicious and the legitimate measurements, then, lead to a delay response $\Delta\tau_{T_{L,cold}}$ equal to 41s (larger than $\Gamma_{T_{L,cold}}^{ref}$) between the threshold exceedance of $S_{T_{L,cold}}^{feed}(t)$ and $S_{T_{L,cold}}^{monitor}(t)$ (see Figure 40(b)), allowing for a (correct) identification of the event as a cyber attack.

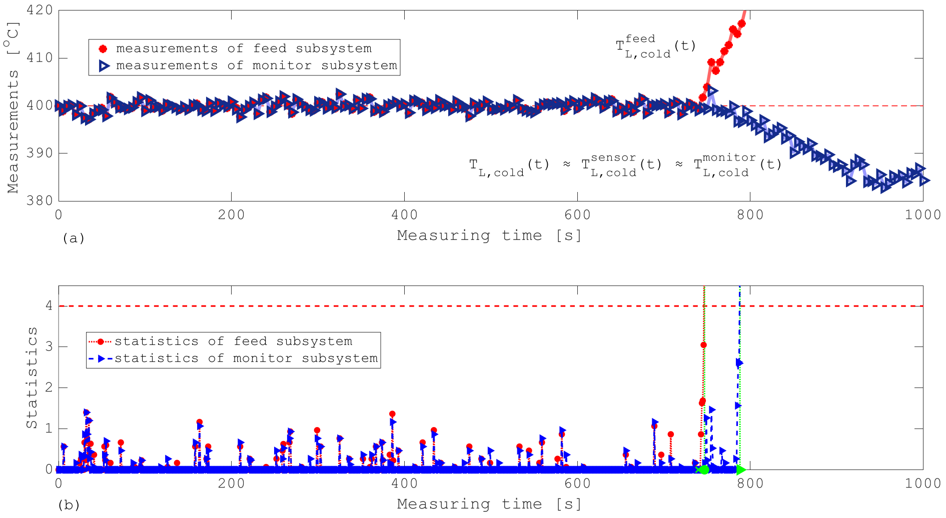


Figure 40 Cyber attack to the computing unit mimicking a drift failure mode: (a) the received measurements of feed and monitor Subsystems in which the cyber attack occurs at time t_R equal to 740s; (b) the corresponding NP-CUSUM statistics for diagnosing the cyber attack

4.2.3.3. Diagnosis of attacks mimicking wider noise failure mode

Figure 41 presents the results of injecting wider noise at time t_R = 750s. This results in a very quick response of both statistics evaluated on the measurements $T_{L,cold}^{feed}(t)$ and $T_{L,cold}^{monitor}(t)$: both statistics reach quickly the threshold $h_{T_{L,cold}}$ (dotted line) and the difference $\Delta\tau_{T_{L,cold}}$ between times to alarm ($\tau_{T_{L,cold}}^{feed}$ and $\tau_{T_{L,cold}}^{monitor}$) turns out to be equal to 0 (i.e., less than $\Gamma_{T_{L,cold}}^{ref}$) (see Figure 41(b)), allowing for a (correct) identification of the event as a sensor failure mode.

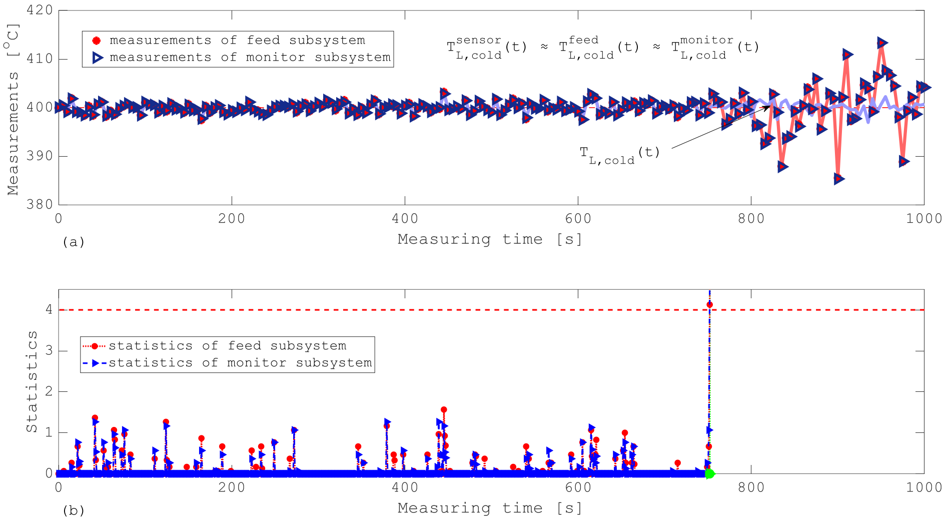


Figure 41 $T_{L,cold}$ sensor wider noise failure mode: (a) the received measurements of feed and monitor Subsystems in which the wider noise failure occurs at time t_R equal to 750s; (b) the corresponding NP-CUSUM statistics for diagnosing the bias failure

Contrarily, Figure 42(a) shows a cyber attack to the computing unit mimicking a wider noise failure mode at t_R=750s, leading $T^{feed}_{L,cold}(t)$ to deviate from the legitimate $T^{monitor}_{L,cold}(t)$. The different values between the malicious and the legitimate measurements, then, lead to a delay response $\Delta\tau_{T_{L,cold}}$ equal to 247s (i.e., larger than $\Gamma^{ref}_{T_{L,cold}}$) at t_M (see Figure 42(b)), allowing for a (correct) identification of the event as a cyber attack.

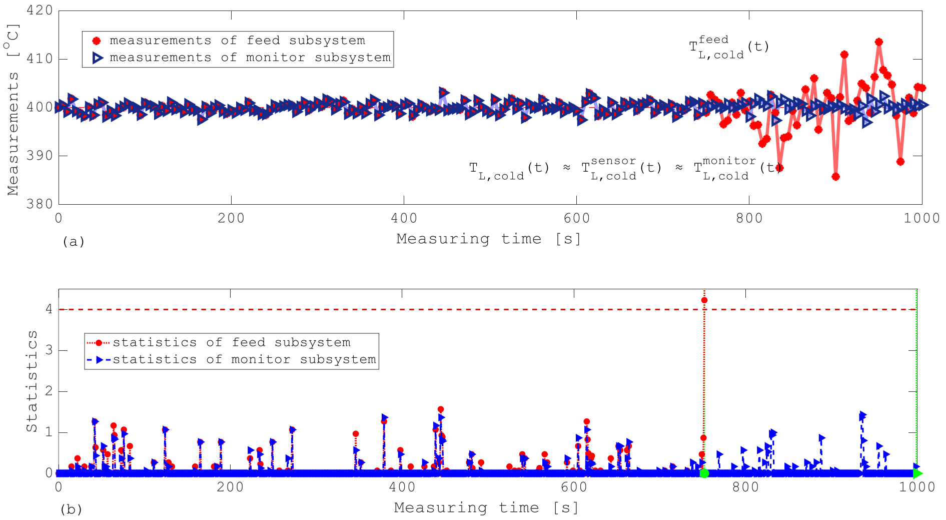


Figure 42 Cyber attack to the computing unit mimicking a wider noise failure mode: (a) the received measurements of feed and monitor Subsystems in which the cyber attack occurs at time t_R equal to 750s; (b) the corresponding NP-CUSUM statistics for diagnosing the cyber attack

4.2.3.4. Diagnosis of attacks mimicking freezing failure mode

Figure 43 presents the results of injecting freezing at time t_R = 460s with the frozen $T_{L,cold}^{sensor}(t)$ equal to 402.53ºC. The freezing results in a very quick response of both statistics evaluated on the measurements $T_{L,cold}^{feed}(t)$ and $T_{L,cold}^{monitor}(t)$: Both statistics reach quickly the threshold $h_{T_{L,cold}}$ (dotted line) and the difference $\Delta\tau_{T_{L,cold}}$ between times to alarm ($\tau_{T_{L,cold}}^{feed}$ and $\tau_{T_{L,cold}}^{monitor}$) turns out to be equal to 0 (i.e., less than $\Gamma_{T_{L,cold}}^{ref}$) (see Figure 43(b)), allowing for a (correct) identification of the event as a sensor failure mode.

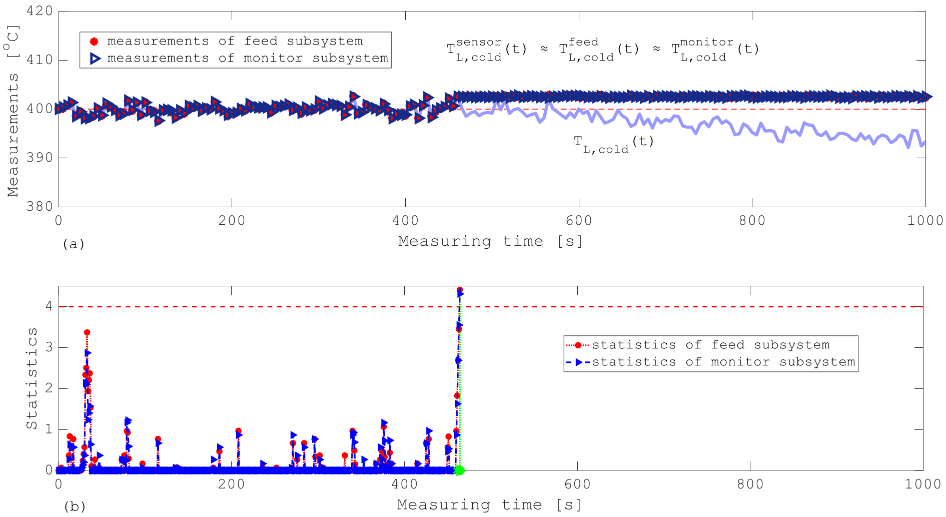


Figure 43 $T_{L,cold}$ sensor freezing failure mode: (a) the received measurements of feed and monitor in which the freezing occurs at time t_R equal to 460s; (b) the corresponding NP-CUSUM statistics for diagnosing the bias failure

Contrarily, Figure 44(a) shows a cyber attack to the computing unit mimicking a freezing failure mode at t_R=460s (with frozen $T_{L,cold}^{sensor}(t)$ again equal to 402.53ºC), leading $T_{L,cold}^{feed}(t)$ to deviate from the legitimate $T_{L,cold}^{monitor}(t)$. The different values between the malicious and the legitimate measurements, then, lead to a delay response $\Delta\tau_{T_{L,cold}}$ equal to 187s (i.e., larger than $\Gamma_{T_{L,cold}}^{ref}$) between the threshold exceedance of $S_{T_{L,cold}}^{feed}(t)$ and $S_{T_{L,cold}}^{monitor}(t)$ (see Figure 44(b)), allowing for a (correct) identification of the event as a cyber attack.

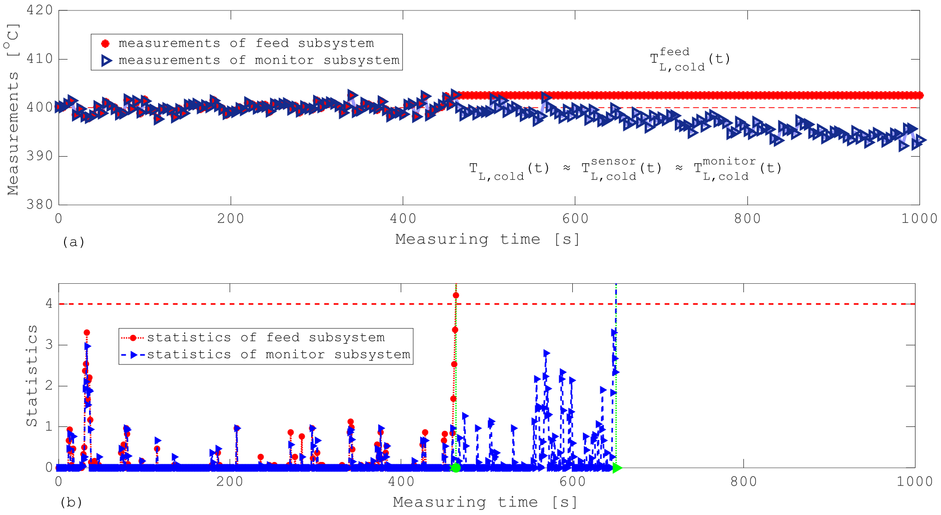


Figure 44 Cyber attack to the computing unit mimicking a freezing failure mode: (a) the received measurements of feed and monitor Subsystems in which the cyber attack occurs at time t_R equal to 460s; (b) the corresponding NP-CUSUM statistics for diagnosing the cyber attack

The results of these illustrative examples show that the NP-CUSUM-based diagnostics approach is capable of diagnosing cyber attacks, distinguishing them from stochastic failures of components, based on the identified rules of assignments.

4.2.4. Performace of the diagnostic approach

The previous examples shown in Section 4.2.3 demonstrate the effectiveness of the NP-CUSUM diagnostics approach. Since the proposed diagnostic approach may suffer from either large false alarm rate (if the threshold is set too small) or high missed alarm rate (if the threshold is set too large) [244], an extensive and massive test with respect to unknown sensor failures and/or unknown cyber attacks is performed for assessing its diagnostic capabilities. We calculate false alarm, missed alarm and misclassification rates with respect to 100 randomly sampled stochastic failures and 100 different cyber attacks for each failure mode (i.e., bias, drift, wider noise or freezing) (thus, a total of N_A=800 runs). At each run of the simulation: a random time t_R within the mission time t_M=1000s and an uncertain parameter value (i.e., b for bias, c for drift, $\delta'(t)$ are sampled from the distributions listed in Table 11 for wider noise or frozen value for freezing) and used to inject a $T_{L,cold}$ sensor failure or a cyber attack to the computing unit. Then, the NP-CUSUM-based diagnostic algorithm is applied to both $T^{feed}_{L,cold}(t)$ and $T^{monitor}_{L,cold}(t)$, to calculate $S_{T_{L,cold}}^{feed}(t)$ and $S_{T_{L,cold}}^{monitor}(t)$, respectively. The diagnostic performances are measured as follows:

- False alarm rate $\alpha_{T_{L,cold}}^{h}$: the probability of either $S_{T_{L,cold}}^{feed}(t)$ or $S_{T_{L,cold}}^{monitor}(t)$ in an accidental scenario exceeding the threshold $h_{T_{L,cold}}$ before t_R.
- Missed alarm rate $\beta_{T_{L,cold}}^{h}$: the probability of neither $S_{T_{L,cold}}^{feed}(t)$ nor $S_{T_{L,cold}}^{monitor}(t)$ in an accidental scenario exceeding the threshold $h_{T_{L,cold}}$ within the mission time t_M.
- Misclassification rate $\gamma\left(\Gamma_{T_{L,cold}}^{ref}\right)$: given a reference delay difference $\Gamma_{T_{L,cold}}^{ref}$, the probability of a misclassified assignment of an event.

Table 31 lists the estimates of $\alpha_{T_{L,cold}}^{h}$ and $\beta_{T_{L,cold}}^{h}$ with respect to the threshold $h_{T_{L,cold}}$ equal to 4.0, among the total of N_A=800 runs of stochastic failures and cyber attacks. The results in the Table show that the total values of $\alpha_{T_{L,cold}}^{h}$ and $\beta_{T_{L,cold}}^{h}$ are equal to 0.0313 and 0.0250, respectively, and the low values are accepted in the diagnostics of cyber attacks of the ALFRED.

Table 31 False and missed alarm rates with respect to $h_{T_{L,cold}}$

Character	Bias	Drift	Wider noise	Freezing	Total
$\alpha_{T_{L,cold}}^{h}$	8/200	8/200	1/200	8/200	25/800=0.0313
$\beta_{T_{L,cold}}^{h}$	13/200	5/200	0/200	2/200	20/800=0.0250

To analyze the effect of an improper choice of $\Gamma_{T_{L,cold}}^{ref}$ that may mistakenly ascribe an accidental scenario to inconsistent reasons and lead to misclassified diagnostics, we estimate $\gamma\left(\Gamma_{T_{L,cold}}^{ref}\right)$ among the N_A=800 scenarios, with respect to different values of $\Gamma_{T_{L,cold}}^{ref}$. We assess the misclassification rates by defining four misclassification types (i.e., Misclassification *A*, *B*, *C* and *D*), that differ in terms of the difference between alarm delays $\Delta\tau_{T_{L,cold}}$ to a reference value $\Gamma_{T_{L,cold}}^{ref}$ in Table 32.

Table 32 Misclassification assessment with respect to $\Gamma^{ref}_{T_{L,cold}}$

Real scenario	Comparison	Assignment	Check	False alarm of	Missed alarm of
Sensor failure	$\Delta\tau_{T_{L,cold}} \leq \Gamma^{ref}_{T_{L,cold}}$	Sensor failure	Correct	-	-
	$\Delta\tau_{T_{L,cold}} > \Gamma^{ref}_{T_{L,cold}}$	Cyber attack	***Misclassification A***	Cyber attack	Sensor failure
	Neither $\tau^{feed}_{T_{L,cold}}$ nor $\tau^{monitor}_{T_{L,cold}}$	Normal condition	***Misclassification B***	-	Sensor failure
Cyber attack	$\Delta\tau_{T_{L,cold}} \leq \Gamma^{ref}_{T_{L,cold}}$	Sensor failure	***Misclassification C***	Sensor failure	Cyber attack
	$\Delta\tau_{T_{L,cold}} > \Gamma^{ref}_{T_{L,cold}}$	Cyber attack	Correct	-	-
	Neither $\tau^{feed}_{T_{L,cold}}$ nor $\tau^{monitor}_{T_{L,cold}}$	Normal condition	***Misclassification D***	-	Cyber attack

Figure 45 shows the calculated misclassification rates $\gamma\left(\Gamma^{ref}_{T_{L,cold}}\right)$ varying with $\Gamma^{ref}_{T_{L,cold}}$ from 0 to 60. $\gamma\left(\Gamma^{ref}_{T_{L,cold}}\right)$ is calculated by summing all the misclassified assignments of the accidental scenarios, which are recorded in the way of false and missed alarm of sensor failures and of cyber attacks, respectively. Results show that the minimum misclassification rate (equal to 0.02875) can be achieved if the categorical difference $\Gamma^{ref}_{T_{L,cold}}$ is optimally equal to 8s or 9s. It is also noted that, the minimum rate being larger than $\beta^{h}_{T_{L,cold}}$ (equal to 0.025) turns out to be reasonable because the identified misclassification scenarios here include the missed alarms identified with respect to $h_{T_{L,cold}}$ equal to 4.0.

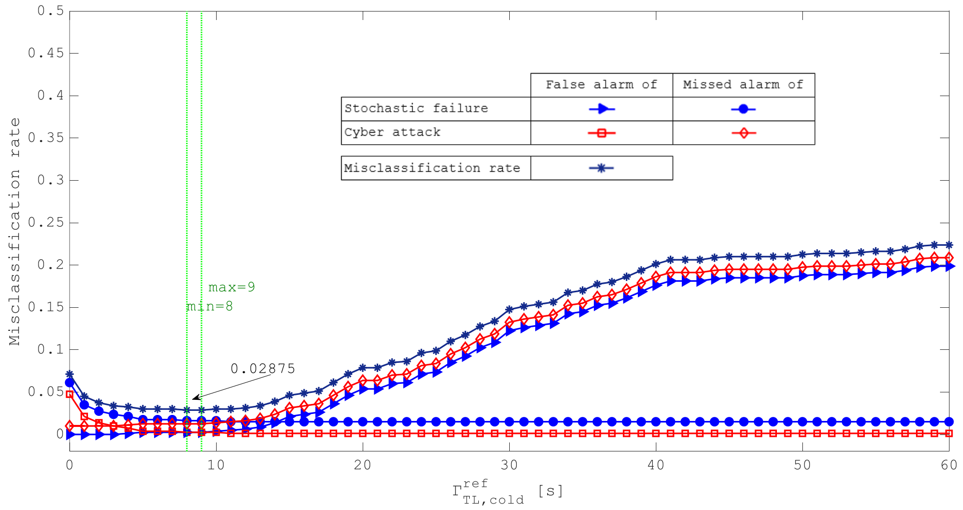


Figure 45 The misclassification rates varying with $\Gamma^{ref}_{T_{L,cold}}$

4.2.5. Reliability assessment of the diagnostic tool under uncertain human operator cognition

Diagnosing the different system abnormal conditions (e.g., component stochastic failures or cyber attacks) and distinguishing become important [53, 122], because would enable the operator to take proper protection decisions and to counteract the effects induced by different disruptive events, and, hence, increase the system safety and security level while reducing the overall corrective maintenance costs [81, 82, 102, 108, 245, 246].

Once informed via HMIs, the operator has to correctly interpret the information conveyed, and translate his assessment in commands actuation for corrective response [247, 248]. The human cognition in monitoring and assessing the information of HMIs can, on one side, improve the performance level of the computerized diagnostic systems embedded within the CPSs when the operator is skilled, whereas, on the other side, can impair the performance level when the cognition is poor [177, 247, 249-251]. In this sense, it is very important to assess the reliability of the online diagnostic system, considering the human role in correctly interpreting the information provided by the HMIs.

A reliability assessment is hereby performed for the NP-CUSUM-based cyber security diagnostic tool, to validate and actualize its capability in diagnosing cyber threats to NPPs. The study takes simultaneously into account two fundamental aspects affecting the reliability assessment: (i) the uncertainty of the NP-CUSUM algorithm, and (ii) the modeling of the uncertainty related to the human operator cognition in interpreting and understanding the outcomes of the diagnostic tool. Human cognition will be modelled by Bayesian Belief Network (BBN) that structures the expert knowledge and understanding on the dependences among human factors (e.g., Performance Shaping Factors (PSFs)) and their causalities to the human cognition errors, in line with [177, 179-182, 252, 253].

4.2.5.1. Online diagnostic performance

Another test with respect to unknown sensor failures and/or unknown cyber attacks is performed for assessing its diagnostic capabilities.

We estimate the correctness of the online assignments of the NP-CUSUM diagnostic tool, with respect to 389 $T_{L,cold}$ sensor failures, 392 DoS attacks and 219 normal operation scenarios (thus, a total of N_v = 1000 tests). At each test scenario j, the NP-CUSUM-based diagnostic algorithm is applied to both $T_{L,cold}^{feed}(t)$ and $T_{L,cold}^{monitor}(t)$ to calculate $S_{T_{L,cold}}^{feed}(t)$ and $S_{T_{L,cold}}^{monitor}(t)$, respectively, with respect to the randomly sampled values of the NP-CUSUM parameters by a MC sampling procedure, allowing for an identification of the event and outputting an assignment i.

Table 33 collects the number of the online indication outputs, and lists the empirical estimates of the correctness rate and the misclassification rate. With respect to the N_v tests, the online diagnostic tool outputs assignments of a (sensor failure), b (DoS attack) and c (normal condition) with 386, 386 and 228 times, respectively. Among them, the probabilities of correct diagnostic turn out to be 0.9611, 0.9819 and 0.8772, respectively, of the different NP-CUSUM online assignments (i.e., a=(component failures), b=(cyber attacks) and c=(normal conditions), when checking with the real scenarios j. It is worth noting that the correct assignment rate of normal conditions being relatively small probably attributes to a fact of the NP-CUSUM algorithm suffering from a relative high missed alarm rate, due to the negligible effects of accidents on the controlled variable or the improper sampling of the uncertain parameters.

Table 33 Performance of the NP-CUSUM diagnostic tool [167]

NP-CUSUM assignment *i*	Occurrence number	Check with real scenario	Probability
Component failure	386	*Correct*	371/386 (0.9611)
		Misclassfication of cyber attack	1/386
		Misclassfication of normal condition	14/386
Cyber attack	386	Misclassfication of component failure	2/386
		Correct	379/386 (0.9819)
		Misclassfication of normal condition	5/386
Normal condition	228	Misclassfication of component failure	16/228
		Misclassfication of cyber attack	12/228
		Correct	200/228 (0.8772)

4.2.5.2. Human operator cognition modeling

The human operator has to judge the correctness of the online diagnostic indications, based on his/her monitoring of the characteristics of physical dynamic processing via HMIs. The human operator may rectify the misclassification of the online diagnostic tool if he/she is well experienced, whereas, may also erroneously respond to a correct online indication and lead to wrong situation assessment when he/she is experiencing stress or depression [254, 255].

4.2.5.2.1. Bayesian Belief Network structuring

We consider a three-phase operator cognitive activity in interpreting and understanding the outcomes of the diagnostic tool, including (1) monitoring/detection (that refers to the operator observing and collecting of the real-time information (e.g., color-coded indications) from the HMIs), (2) situation assessment (that refers to the operator developing and updating his/her mental representation of the specific current situation) and (3) response planning (that refers to the operator diagnostic decision-making for further response plan to the current situation) [180, 247, 254, 256, 257].

Facing to an online diagnostic outcome, the operator develops his/her cognition relying on both the instantaneous understanding of the specific system situation and the mental cognition built up through formal education, system-specific training, and operational experience, namely, the knowledge base [258]. Particularly, the operator simultaneous understanding of the real-time system observations affects his/her performance at all the three phases, whereas the mental cognition responding to the specific diagnostic outcome affects his/her performance at phases (2) and (3). Besides, a severe system situation potentially makes the operator more stressful and, eventually, impede the operator from completing the diagnostic task [180, 253]. Thus, the human operator cognition in interpreting the online diagnostic outcomes is mutually affected by the system situation level, the human mental level and the human stress level, as sketched in Figure 46.

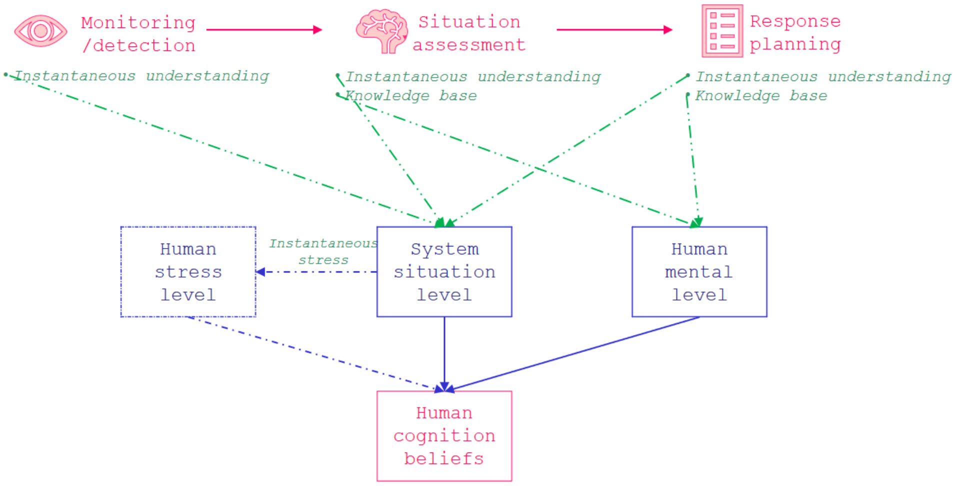


Figure 46 The operator cognitive activity in diagnosing anomalies

We, then, identify the Performance Shaping Factors (PSFs) and their dependences involved in the human mental model, the system situation model and the human stress model. Table 34 lists the identified PSFs with states and descriptions, and Figure 47 develops the relationships of PSFs (i.e., parent nodes) affecting the operator cognitive activity (i.e., child node), namely, a BBN model elicited by expert judgment [178, 181].

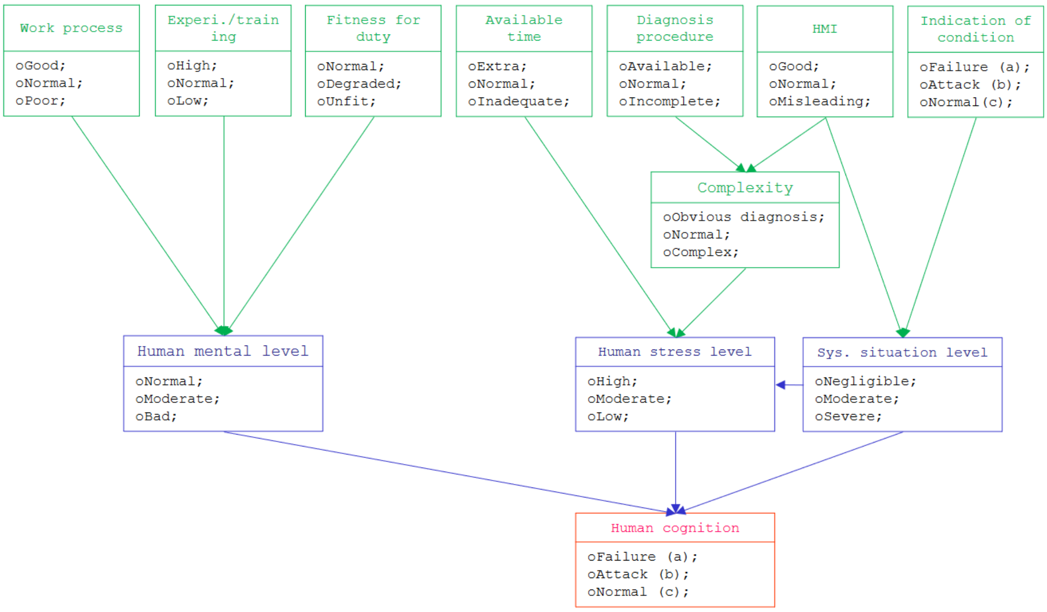


Figure 47 The BBN model of the human operator diagnostic cognition

Table 34 Identification of PSFs affecting the human diagnostic cognition

Child node, n_c^{β} with states $S_c^{\beta,\gamma}$, β=			Parent nodes, n_p^{α}, α=	States, $S_p^{\alpha,\gamma}$	Descriptions

(1) Human cognition beliefs	(2) Human mental level		(1) Work process	Good;
				Normal;
				Poor.
					The way to diagnose anomalies, e.g., coordination and communication between operators, management support, strategy handling given situations, and corrective action programs, etc. [259-261].

			(2) Diagnosis experience/ training	High;
				Normal;
				Low.
					The operator knowledge base, experience and training involved in the diagnostic task [261].

			(3) Fitness of duty	Normal;
				Degraded;
				Unfit.
					The operator physical and mental fitness to perform the diagnosis task at the time [259, 261].

	(3) System situation level	/	(4) Available diagnosis time	Extra;
				Normal;
				Inadequate.
					The operator's available time to diagnose an abnormal event [261].

		(5) Diagnosis complexity	(5) Diagnosis procedure	Available;
				Normal;
				Incomplete.
					The existence of feasible procedures for the diagnosis and response planning tasks [261, 262].

			(6) HMI	Good;
				Normal;
				Misleading.
					The availability of real-time physical information from HMIs for the operator to carry out the diagnostic task [261].

	(4) Human stress level		(6) HMI		
			(7) Indication of condition	i=a; i=b; i=c.	The obviousness of the online indications to help the operator diagnose the anomaly in real-time monitoring [263].

Note:

α= 1, 2, 3, 4, 5, 6 or 7 for parent nodes;
β = 1, 2, 3, 4 or 5 for child nodes;
γ = 1, 2 or 3 for all the nodes;

4.2.5.2.2. Quantitative analysis considering uncertainty

As shown in Table 34 and Figure 47, the operator diagnostic cognition depends both on the understanding of an online output of the diagnostic tool *i* (= sensor failures (a), DoS attacks (b) or normal conditions/missed alarms (c)) and on the knowledge base and experience towards a class of real events *j* (=a, b or c). Thus, different combinations of real scenarios and online indications, i.e., (*j*,*i*), to different extent, change the operator's attention of monitoring and, then, affect the correctness of the anomaly diagnostic.

Given any an online output *i* represented on the HMI, only a consistent operator decision *k* with the real scenario *j* can allow for a correct diagnostic of the event. The probability of correct diagnostic can, thus, be expressed:

$$p_{\mathrm{correct}}^{i} = \sum_{j=k=a}^{c} p\left(j, k = j \mid i\right) \tag{4-40}$$

where *i*=a, b or c, and $p(j, k = j|i)$ is the probability of the operator decision *k* being consistent with the real scenario *j*, conditional on the indication *i*. According to the chain rule of conditional probability, Eq. (4-40) can change to:

$$p_{\mathrm{correct}}^{i} = \sum_{j=k=a}^{c} p\left(k = j \mid j, i\right) \cdot p\left(j \mid i\right) \tag{4-41}$$

where $p(k = j|j, i)$ is the probability of the operator correctly recognizing the accidental event *j*, conditional on the indication *i*, and is dependent on the operator cognitive ability in responding to undesired accidental events [180, 257]; whereas, $p(j|i)$ is the probability of the real scenario being *j*, conditional on the indication *i*, as listed in Table 33.

The BBN model of Figure 47 can represent an operator's instantaneous understanding of an evidence *i* (i.e., *p*(*i*)=1, *i*=a, b or c) and his/her experience and skills responding to the accidental events *j*, and can output the estimates of $p(k = j|j, i)$ conditional on the combination (*j*,*i*), provided the Conditional Probability Distributions (CPDs) of the parent nodes (i.e., PSFs) and the Conditional Probability Tables (CPTs) at the child nodes are known.

In order to treat the uncertainty on the PSFs with their dependences and causalities to the human cognition, we illustrate the use of a functional interpolation method [182] (shadowed in

Figure 48) embedded within a MC simulation, for populating the CPTs at the child nodes that is fundamental for providing us with the estimates of $p(k=j|j,i)$ conditional on the combination (j,i), and of human correct interpretation probability $p^{i}_{correct}$ with respect to NP-CUSUM online assignments.

At the *m*-th MC run, m=1, 2, ..., N_m, of Figure 48:

(1) Orderly set j = a, b and c, and at the j-th selection, invoke the relative distributions of PSFs (see Appendix B) indicating the operator knowledge base and experience to the class of events j;

(2) Orderly set i = a, b and c, and at the i-th selection, set the evidence of the parent node n_p^7 (indication of condition) as i, i.e., $p\left(s_p^{7,\gamma}=i\right)=1$ and $p\left(s_p^{7,\gamma}\neq i\right)=0$;

(3) Sample the CPDs (i.e., $p_m\left(s_p^{\alpha,\gamma}\right)$, the conditional probability of the states $s_p^{\alpha,\gamma}$) of the parent nodes n_p^{α}, α = 1, 2, 3, 4, 5, 6, from the related distributions;

(4) Populate the CPTs of the child nodes n_c^{β} by illustrating the use of the five-step functional interpolation method:

(4a) Sample the mean and standard deviation values ($u\left(e_c^{\beta,\theta}\right)$ or/and $\sigma\left(e_c^{\beta,\theta}\right)$, where $e_c^{\beta,\theta=\text{anchor}}$ are the selected anchors at the anchor CPT of the child node n_c^{β} , from the expert-judged distributions (see Appendix C);

(4b) Linearly interpolate the missing mean and standard deviation values ($u\left(e_c^{\beta,\text{other}}\right)$ or/and $\sigma\left(e_c^{\beta,\text{other}}\right)$) of the other elements at the anchor CPT of n_c^{β} and; then,

(4c) Formulize all the elements of each anchor CPTs to the corresponding uniform distributions $N\left(u\left(e_c^{\beta,\theta}\right),\sigma\left(e_c^{\beta,\theta}\right)\right)$, which are defined on the underlying rule that the pdf values at 1, 2 (and 3) represent the CPD scales of the states $s_c^{\beta,\gamma}$ at the θ-th element of the child node n_c^{β} CPT (see Appendix C);

(4d) With respect to each element, normalize the pdf scales sum to 1, being the CPD of the θ-th element;

(4e) Collect the CPDs and, build the CPTs for each child node;

(5) Quantify the BBN model with the sampled CPDs of parent nodes and CPTs of child nodes, and estimate the operator correct diagnostic probability $p_m\left(k=j\mid j,i\right)$;

(6) At the *m*-th MC run, collect the estimates of $p_m\left(k=j\mid j,i\right)$ of all nine combinations (*j*,*i*);

(7) Feed $p\left(k=j\mid j,i\right)$ and the tested $p\left(j\mid i\right)$ values to Eq. (4), to obtain the estimates of the correct diagnostic probabilities $p^i_{\text{correct},m}$, with respect to different online indications i.

(8) Repeat steps (1) to (7) for N_m times, and obtain the confidence interval of the p^i_{correct}, with respect to different online indications *i*;

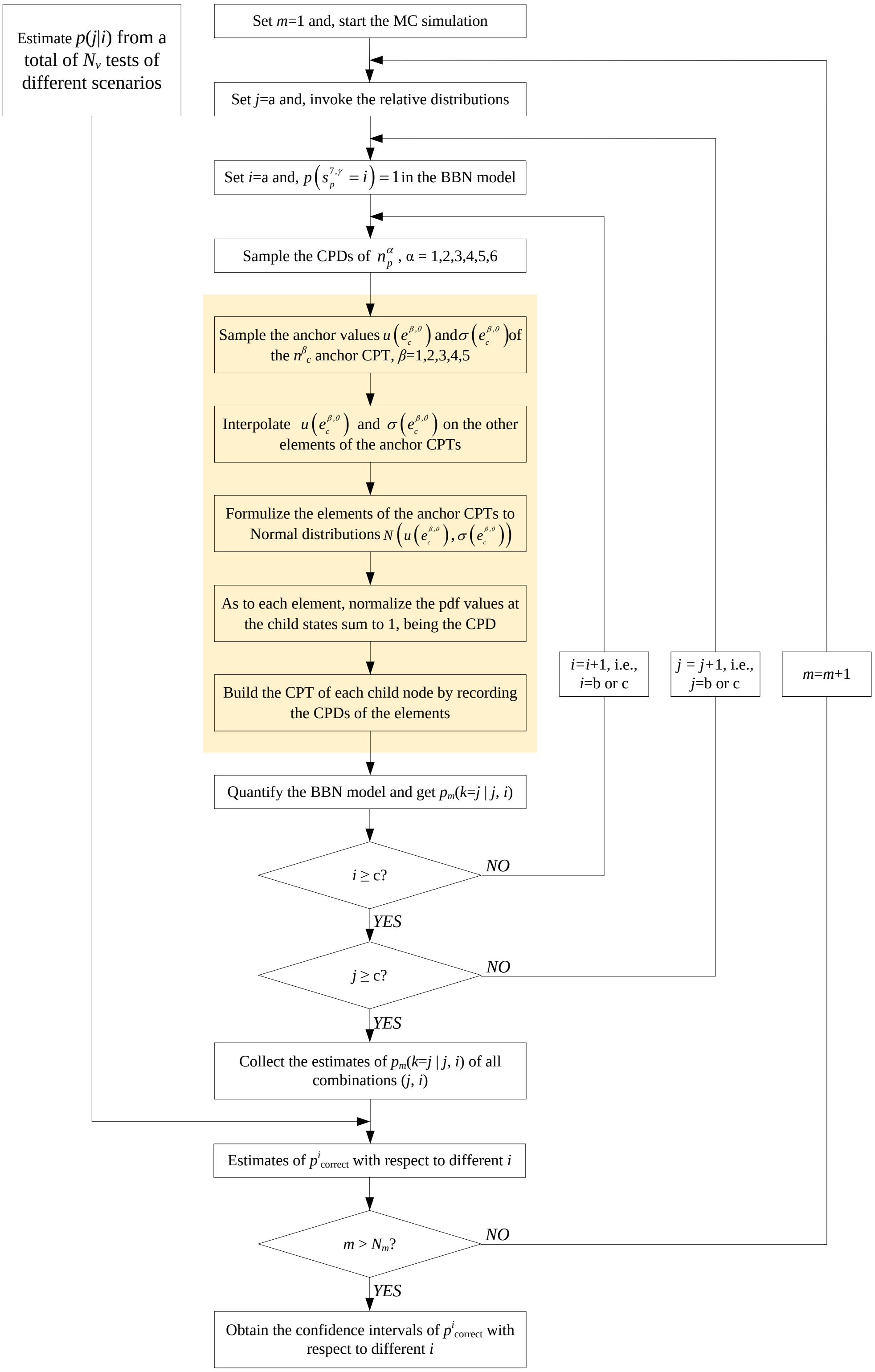

Estimate $p(j|i)$ from a total of N_v tests of different scenarios
Set m=1 and, start the MC simulation
Set j=a and, invoke the relative distributions
Set i=a and, $p\left(s_p^{7,\gamma}=i\right)=1$ in the BBN model
Sample the CPDs of n_p^{α}, α = 1,2,3,4,5,6
Sample the anchor values $u\left(e_c^{\beta,\theta}\right)$ and $\sigma\left(e_c^{\beta,\theta}\right)$ of the n^{β}_c anchor CPT, β=1,2,3,4,5
Interpolate $u\left(e_c^{\beta,\theta}\right)$ and $\sigma\left(e_c^{\beta,\theta}\right)$ on the other elements of the anchor CPTs
Formulize the elements of the anchor CPTs to Normal distributions $N\left(u\left(e_c^{\beta,\theta}\right),\sigma\left(e_c^{\beta,\theta}\right)\right)$
As to each element, normalize the pdf values at the child states sum to 1, being the CPD
Build the CPT of each child node by recording the CPDs of the elements
Quantify the BBN model and get $p_m(k=j \mid j, i)$
$i \geq$ c?
NO
YES
$j \geq$ c?
NO
YES
Collect the estimates of $p_m(k=j \mid j, i)$ of all combinations (j, i)
Estimates of p^i_{correct} with respect to different i
$m > N_m$?
NO
YES
Obtain the confidence intervals of p^i_{correct} with respect to different i
i=i+1, i.e., i=b or c
j = j+1, i.e., j=b or c
m=m+1

Figure 48 The flowchart for estimating the correct diagnostic probability

4.2.4.2.3. Results

Figure 49 shows the assessment results for the accidental scenarios that have been diagnosed and assigned by the NP-CUSUM diagnostic tool (as results are shown in Table 33).

Considering the operator cognitive activity in interpreting the online diagnostic outputs, Figure 49 results in the double-sided 50% confidence intervals (shadowed boxes in Figure 49) of the correct diagnostic probabilities, with mean values equal to 0.9662, 0.9230 and 0.9429, respectively, of the different NP-CUSUM online assignments (i.e., a=(component failures), b=(cyber attacks) and c=(normal conditions), respectively). Comparing with the correct probability of online diagnostic equal to 0.9611, 0.9819 and 0.8772 (Table 33), the human operators can effectively increase the reliability in diagnosing components failures and false negatives, but perform worse on the less predictable cyber attacks.

Thereby considering an experienced operator, the diagnostic performance improves with the mean values becoming 0.9666, 0.9881 and 0.9561, respectively, of the different NP-CUSUM assignments (i.e., a, b and c, respectively). The narrower confidence interval (colorless boxes in Figure 49) and increased mean value (equal to 0.9881) of the estimates of correct diagnostic probability, particularly with respect to b, suggest the necessity of improving the operators experience and skills in dealing with unforeseeable cyber attack events.

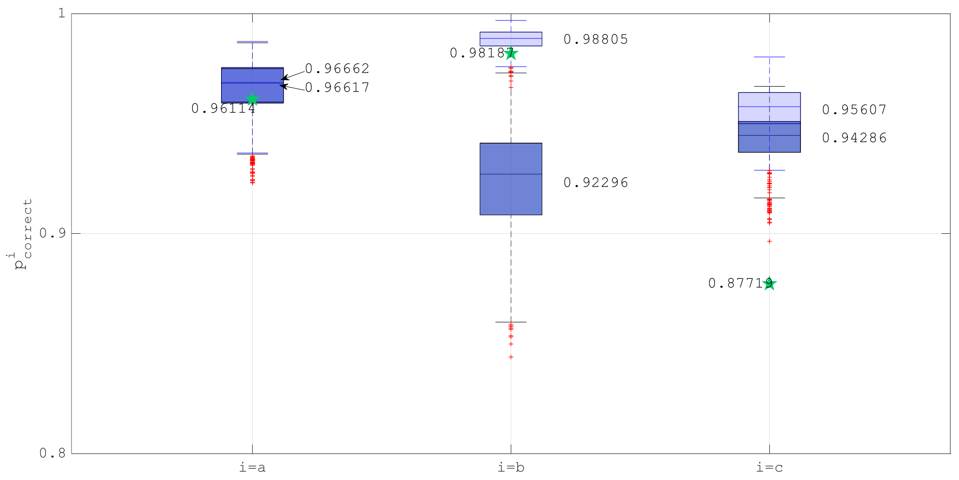


Figure 49 Estimates of the confidence intervals of the correct diagnostic probabilities (i=a refers to online assignment as component failures, i=b to online assignment as cyber attacks, and i=c to online assignment as normal conditions) [167]

4.2.4. Conclusions

In this chapter, we have presented a NP-CUSUM approach to enable real-time diagnosis of cyber attacks on CPSs. The diagnostics approach allows distinguishing between components failures and cyber attacks to the controllers, guiding decisions for recovering CPSs from anomalies.

The diagnostic performance of the approach has been analyzed by the false and missed alarm rates, with reference to a prespecified threshold, and the misclassification rates varying with the reference delay differences for identifying a cyber attack or a sensor failure.

We have applied the diagnostics approach to the digital I&C system of the ALFRED. Cyber breach events attacking the embedded CPS controllers and sensor failures are injected by a MC sampling procedure, at random times and with random magnitudes. Results show that the diagnostic approach is capable of identifying most of the generated failure/attack scenarios, with low false alarm rate, missed alarm rate and misclassification rate (Paper 1[B]).

To verify the diagnostic capability of the NP-CUSUM online diagnostics tool while accounting for the human operator cognition in interpreting the online diagnostic outputs, an assessment has been performed by a MC simulation. The BBN approach has been used for modeling the human cognitive activities (Paper 5[J]). The human operators can effectively increase the reliability in diagnosing components failures and false negatives, but perform worse on the less predictable cyber attacks. Thereby considering an experienced operator, the narrower confidence interval and increased mean value of the correct diagnostic probability, particularly with respect to cyber attacks, suggest the necessity of improving the operators experience and skills in dealing with unforeseeable cyber attack events (Paper 5[J]).

5. CONCLUSIONS AND FUTURE PERSPECTIVES

Cyber-Physical Systems (CPSs) must perform safely and securely in their functions to enable opportunities of connectivity, real-time monitoring, communication, dynamic control and decision support during normal operation of industrial systems, as well as in case of accidents.

However, CPSs are subjected to degradations and failures of their physical components, and to intentional or accidental breaches in the cyber components. Thus, CPSs failure analysis must comprise safety and security aspects.

In the Ph.D. thesis presented, a modelling and simulation framework for the analysis of failures of CPSs has been developed, considering both safety and security aspects, for I. identification and prioritization of hazards and threats (to identify the conditions that trigger anomalies in the systems and their causes), II. failure scenarios modeling and simulation (to characterize the system behavior under different operational conditions, including hazardous and malicious ones), III. consequence analysis (to explore the effects of stochastic component failures and cyber attacks onto the CPS functionality) and, IV. protection design (to take decisions on recovery measures for increasing system resilience).

The proposed framework is expected to provide results that help the analysts to identify hazards and threats of CPSs, analyze their causes, model their potential scenarios and consequences, and propose decisions for system protection and resilience.

Application of the framework has concerned the typical Reactor Protection System (RPS) of Nuclear Power Plants (NPPs) and the digital I&C system of an Advanced Lead-cooled Fast Reactor European Demonstrator (ALFRED)). A Sensitivity Analysis (SA) has been performed to identify the components of a system that most deserve accurate modeling of aging- and degradation-dependent transition rates, for accurate system reliability assessment and for trading model accuracy and computational demand for practical reliability assessment based on Multi-State Physics Modeling (MSPM). The SA performed on the RPS of NPPs based on moment-independent sensitivity measures, such as Hellinger distance and Kullback-Leibler divergence, has led to the focus on the accurate modeling of a Resistance Temperature Detector (RTD) for system reliability assessment. A three-loop Monte Carlo (MC) simulation scheme has been developed to operationalize the MSPM approach for large scale systems, and to quantify and control the uncertainty affecting the system reliability model.

A MC-based modelling and simulation framework has been developed for generating cyber attack scenarios in CPSs and accounting for multiple failure modes of attacked components of the CPSs, to test the effects of the cyber threats on the system functionality and integrity, and to prioritize the most vulnerable components for cyber security protection decision-making. A safety margin estimation approach has been proposed for cyber threat prioritization. Safety margins of the safety parameters are estimated by a Bracketing Order Statistics (OS) approach, with respect to the one- and two-sided thresholds. The results of the case study, i.e., the digital I&C system of the ALFRED, identify actuators as the most vulnerable CPS components, their failures leading more easily to the loss of system functionality and integrity, along with the lead temperature sensor, which is relevant component for the control of the temperature lead in the cold pool.

In modeling and simulation for protection design, on one side, an Adversarial Risk Analysis (ARA) approach has been proposed for analyzing decisions between intelligent adversaries providing a novel one-sided (i.e., defender) prescriptive support strategy for optimizing the defensive resource allocations based on a subjective expected utility model. A MC approach has been embedded into the ARA model for treating uncertainties in the decisions of the adversaries, for improving confidence in obtaining the optimal defense resource allocation, leveraging robustness of protection actions on the CPS with uncertain malicious threats.

On the other side, a Non-Parametric Cumulative Sum (NP-CUSUM) approach has been presented to enable real-time diagnosis of cyber attacks on CPSs. The diagnostics approach has been demonstrated to be with very low false and missed alarm rates, with reference to a prespecified threshold, and very low misclassification rates varying with the reference delay differences, when applied to the digital I&C system of the ALFRED. A reliability assessment of the NP-CUSUM-based online diagnostic tool has demonstrated that human operators can effectively increase the reliability in diagnosing components failures, but suggested the necessity of improving the operators experience and skills in dealing with cyber attack events.

5.1. Original contributions of the Ph.D. work

The main original contributions of the research lie in the methodological developments done and presented here for treating both safety and security aspects of CPSs. In particular, novel methods have been developed for addressing the following research issues: 1) giving due account to uncertainties affecting aging, degradation and stochastic failures of CPS components, 2) giving due account to the uncertainties that affect threats and vulnerabilities of CPS to unexpected malicious external attacks, and 3) protection design of CPSs giving due account to cyber attacks and component stochastic failures.

The proposed framework and the practical contributions with respect to the research objectives are summarized in Table 35, compared with the current the state of the art approaches.

Table 35 Original contributions of the Ph.D. work

	Objectives	Hazard analysis		Threat analysis	
		Methods	Results	Methods	Results
I	Identification and prioritization of hazards and threats	Expert judgment-based approaches	• Overall gathering of component failure modes has been achieved mainly based on the analyst experience and brainstorming activity.	Expert judgment-based approaches	• Overall gathering of vulnerabilities has been achieved mainly based on the analyst experience and brainstorming activity.
		I.1 Sensitivity Analysis (SA)	• The SA provides of the indication to the analysis of which components deserve more accurate modeling, according to their contribution to the system reliability.	**I.2** MC-based exploration framework	• The approach generate and process cyber attack scenarios in CPSs, for accounting for multiple failure modes of attacked components of the CPSs.
II	Failure scenarios modeling and simulation	Static/dynamic graphic model (e.g., MCM, FTA, ETA)	• Binary-state modeling approachs neglected the impacts of physical knowledge that accounts for the components aging, degradation and stochastic failure processes on the system reliability assessment.	Conceptual or numerical model (e.g., attack tree)	• The models can understand threats to physical systems, but missed the attacker's interests in injecting all possible failures to CPSs.
		II.1 Multi-State Physics Modeling (MSPM)	• Reliance on physical knowledge accounting for aging and degradation process, MSPM provides a realistic representation of the CPS component degradation progression.	**II.2** MC-based exploration framework	• The approach simulate the effects of all possible attacks aiming at damaging different components of the CPSs, generating different scenarios in the physical domain which lead to different consequences.
III	Consequence analysis	Conceptual, simulation or numerical methods	• The methods can understand physical phenomena that the components failures lead to but, fail to control and quantify the uncertainty affecting the system reliability.	Conceptual, simulation or numerical methods	• The methods can understand physical phenomena that the cyber attacks scenarios lead to but, fail to control and quantify the uncertainty in system responses to cyber threats.
		III.1 Three-loop Monte Carlo (MC) simulation	• Overestimation of the system unreliability is reduced, especially at the early stage of the system life; • The narrower confidence interval of the system unreliability of the RPS-MSPM with respect to the RPS-MCM would more likely induce the decision-maker to rely on the reliability assessment measures provided by the MSPM; The approach allows balancing modeling efforts and computational demand with accuracy of the results.	**III.2** Safety margins estimation approach	• The approach prioritize the components most vulnerable to cyber attacks to CPSs, for guiding cyber security protection decision-making.

IV	Protection design	Traditional game theory	• The models are performed from the viewpoint of a neutral opponent governing the attack/defense loss, under the strong assumptions of mutually consistent knowledge between defender and attacker; • The solution is relatively less realistic and, in practice assumes the maximum quantity of defense resources to be installed with the maximum allowed budget for defense.
		IV.1 Adversarial Risk Analysis (ARA) approach	• The ARA provide a novel one-sided (i.e., defender) prescriptive support strategy for optimizing allocation of resources for the defensive barriers based on a subjective expected utility model; • The optimal decision reaches the one-sided prescriptive optimal decision against all possible uncertain cyber attacks without reaching the maximum budget.
		Data-based approach	• The approaches can prompt recognize the predefined components stochastic failures or cyber attacks, but are not capable of distinguishing cyber attacks from component stochastic failures in CPSs.
		IV.2 Non-Parametric Cumulative SUM (NP-CUSUM) approach	• The NP-CUSUM online diagnostic approach can promptly distinguish cyber attacks from component failures in CPSs, for guiding decisions for the CPSs recovery from anomalous conditions; • The diagnostic performance has been demonstrated with very low false and missed alarm rates with reference to a prespecified threshold, and very low misclassification rates varying with the reference delay differences for identifying a cyber attack or a sensor failure.

.

5.2. Future perspectives

Under the proposed modeling and simulation framework for the analysis of CPSs failures comprising both safety and security aspects, future work can be devoted to the development of a general ARA framework for defense resource allocation capable of accounting for the reduction of CPS security with the stochastic degradation progression of the components/subsystems of the CPSs and allowing to optimize both the defensive resource allocation to cyber attacks and the maintenance strategy for coping with component degradations.

APPENDIX A: THE NON-PARAMETRIC CUMULATIVE SUM ALGORITHM

Without loss of generality, let us consider an accidental scenario *a* simulated over a mission time t_M, during which a cyber attack occurs at random time t_R ($t_R<t_M$). Considering a time interval *dt*, we can define the pre-attack signal mean value $\mu_Y\big(Y(t)\big) = \sum_t Y(t)/t$, $t = dt, 2dt, \ldots, t$, ($t<t_R$), where $Y(t)$ is the measurement Y of a controlled variable y at time t under normal operation conditions (see Figure 50(a), for example). Assume that DoS attacks lead to arbitrary and abrupt changes in the distributions of observations, such that the (unknown) post-attack mean value results to be $\theta_Y\big(Y(t)\big) = \sum_t Y(t)/(t-t_R)$, $t = t_R, t_R+dt, t_R+2dt, \ldots$.

We define a score function $g_Y(Y(t))$ as:

$$g_Y\left(Y(t)\right)=\sum_t \omega_y \cdot \Lambda\left(Y(t)\right)=\sum_t \omega_y \cdot \left(\left|Y(t)-\mu_Y\right|-c_y(t)\right) \qquad \text{(A-1)}$$

where ω_y is a positive weight that is used for normalizing $\Lambda\big(Y(t)\big)$ and chosen equal to $1/\sigma_Y$, where σ_Y is the standard deviation of $Y(t)$, $t = dt, 2dt, \ldots$, and the parameter $c_y(t)$ depends on the past t-1 measurements as in Eq. (A-2):

$$c_y(t)=\varepsilon_y \cdot \hat{\theta}_Y(t) \qquad \text{(A-2)}$$

where ε_y is a tuning parameter belonging to the interval (0,1) and $\hat{\theta}_Y(t)$ is an estimate of the unknown mean value $\theta_Y(Y(t))$. In practice, it is difficult to estimate $\hat{\theta}_Y(t)$ on-line. Hence, Eq. (A-1) is simplified in:

$$\Delta g_Y\left(Y(t)\right)=\omega_y \cdot \left(\left|Y(t)-\mu_Y\right|-c_y\right) \qquad \text{(A-3)}$$

The score function $S_Y(t)$ adopted in the NP-CUSUM algorithm is, then, defined as:

$$S_Y(t)=\max\left\{0, S_Y(t-1)+\Delta g_Y\left(Y(t)\right)\right\} \qquad \text{(A-4)}$$

where, $S_Y(0) = 0$.

In practice, with respect to a stream of measurement $Y(t)$, the NP-CUSUM statistics $S_Y(t)$ remain close to zero or slightly positive under normal operation conditions, whereas, it starts drifting and increasing when a cyber attack occurs at time t_R and, ends up with exceeding a predefined positive threshold h_y (see Figure 50(b)). An alarm can be triggered when $S_Y(t)$ reaches h_y at the time of alarm:

$$\tau_Y = \min\left\{t \geq 1 : S_Y(t) \geq h_y\right\} \tag{A-5}$$

The detection delay $d\tau_Y$ between t_R and τ_Y depends on the choice of h_y. A good diagnostic algorithm is expected to perform with a low False Alarm Rate (FAR) and a small value $d\tau_Y$.

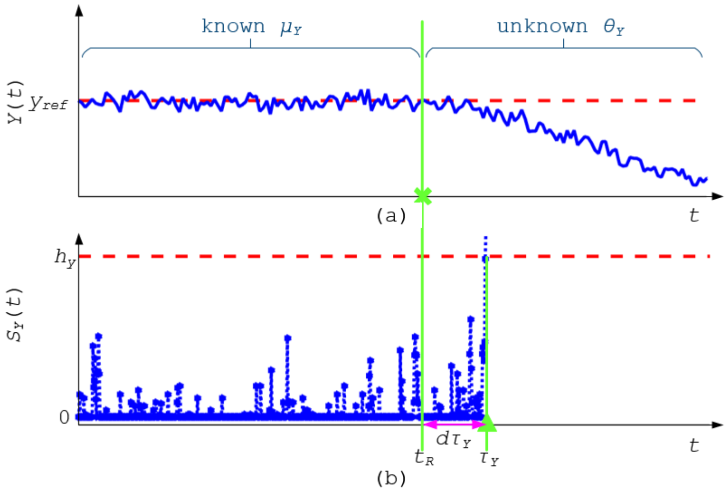


Figure 50 The NP-CUSUM algorithm: (a) a stream of measurement Y(t) of an accidental scenario in which a cyber attack occurring at time t_R; (b) the corresponding NP-CUSUM statistic $S_Y(t)$ for diagnosing the cyber attack at the time to alarm τ_Y

APPENDIX B: DISTRIBUTIONS OF PERFORMANCE SHAPING FACTORS

With respect to the sensor failures and missed alarms conditions (j = a or c), the expert elicits the probability for the PSFs states distributed as uniform distributions (see Table 36), based on the empirical values recommended in [264]; whereas, the expert concerns that the operators are commonly inexperienced and the diagnosis procedures are relatively incomplete with respect to the DoS attacks events (j = b), and, thus, redefine the probability distributions for the states of n_p^2 (Diagnosis experience/ training) and n_p^5 (Diagnosis experience/ training), listed in the last column of Table 36.

At the m-th run of the MC simulation for the quantitative analysis of the BBN model, with respect to a PSF n_p^α, we sample three values from the probability distributions of the PSF states $S_p^{\alpha,\gamma}$, γ = 1, 2 and 3, respectively, and normalize the values sum to 1, being the CPD of $S_p^{\alpha,\gamma}$.

Table 36 Identification of probability distributions for the states of PSFs

Parent nodes, n_p^α, α=	States, $S_p^{\alpha,\gamma}$, γ=	Distributions under j = a and c	Distributions under j = b
(1) Work process	(1) Good	U[0.70, 1.00]	same as Column 3
	(2) Normal	U[0.00, 0.30]	same as Column 3
	(3) Poor	U[0.00, 0.10]	same as Column 3
(2) Diagnosis experience/ training	(1) High	U[0.30, 0.60]	U[0.00, 0.20]
	(2) Normal	U[0.20, 0.50]	U[0.20, 0.50]
	(3) Low	U[0.00, 0.30]	U[0.50, 0.80]
(3) Fitness of duty	(1) Normal	U[0.10, 0.25]	same as Column 3
	(2) Degraded	U[0.70, 1.00]	same as Column 3
	(3) Unfit	U[0.00, 0.10]	same as Column 3
(4) Available diagnosis time	(1) Extra	U[0.10, 0.30]	same as Column 3
	(2) Normal	U[0.50, 0.80]	same as Column 3
	(3) Inadequate	U[0.00, 0.25]	same as Column 3
(5) Diagnosis procedure	(1) Available	U[0.30, 0.70]	U[0.10, 0.30]
	(2) Normal	U[0.20, 0.40]	U[0.20, 0.40]
	(3) Incomplete	U[0.00, 0.40]	U[0.50, 0.70]
(6) HMI	(1) Good;	U[0.70, 1.00]	same as Column 3
	(2) Normal;	U[0.10, 0.25]	same as Column 3
	(3) Misleading.	U[0.00, 0.05]	same as Column 3

APPENDIX C: ANCHOR CONDITIONAL PROBABILITY TABLES OF THE CHILD NODES

As suggested in [182], we build the anchor CPTs for the child nodes n_c^{β} (β = 1, 2, 3, 4, 5) of the BBN model of Figure 47, as listed in Tables 37, 38, 39, 40 and 41, respectively. In each Table, the anchor elements are shaded with the expert-judged values or/and distributions of the means and standard deviations (i.e., $u\left(e_c^{\beta,\theta=\text{anchor}}\right)$ or/and $\sigma\left(e_c^{\beta,\theta=\text{anchor}}\right)$. It is noticed that the states of the child nodes $S_c^{\beta,\gamma}$ are assigned with the values 1, 2 (and 3) for identifying the corresponding CPD scales (i.e., pdf values at 1, 2 (and 3)), once the uniform distributions at all the elements of the anchor CPTs are generated.

Table 37 The anchor CPT of n_c^1 (Human cognition beliefs)

Human mental level		Normal			Moderate			Bad		
Human stress level		Low	Moderate	High	Low	Moderate	High	Low	Moderate	High
System situation level	Negligible	1.00; U[0.20,0.30]		U[1.20,1.50]; U[0.20,0.40]				U[1.20,1.50]; U[0.20,0.25]		2.00; U[0.50,0.70]
	Moderate									
	Severe	1.00; U[0.20,0.40]		U[1.20,1.50]; U[0.20,0.50]				U[1.20,1.50]; U[0.50,0.70]		2.00; U[0.70,1.00]

Note:

1) In each shaded anchor element, the first value/distribution refers to the mean value/distribution and, the second one refers to the standard deviation value/distribution;
2) The n_c^1 states correct (i.e., *k=j*) and incorrect (i.e., *k≠j*) diagnostic are assigned with the values 1 and 2, respectively.

Table 38 The anchor CPT of n_c^2 (Human mental level)

Work process		Good			Normal			Poor		
Experience/training		High	Normal	Low	High	Normal	Low	High	Normal	Low
Fitness of duty	Normal	1.00; U[0.20,0.30]		2.00; U[0.60,0.80]				1.00; U[0.40,0.70]		2.00; U[0.20,0.40]
	Degraded									
	Unfit	1.00; U[0.20,0.50]		2.00; U[0.60,0.90]				1.00; U[0.40,0.70]		3.00; U[0.20,0.50]

Note:

1) In each shaded anchor element, the value refers to the mean value and, the distribution refers to the standard deviation distribution;
2) The n_c^2 states Normal, Moderate and Bad are assigned with the values 1, 2 and 3, respectively.

Table 39 The anchor CPT of n_c^3 (Human stress level)

Available time		Extra			Normal			Inadequate		
Diagnosis complexity		Obvious	Normal	Complex	Obvious	Normal	Complex	Obvious	Normal	Complex
System situation level	Negligible	1.00; U[0.20,0.30]		2.00; U[0.20,0.40]				1.00; U[0.50,0.80]		3.00; U[0.70,1.00]
	Moderate									
	Severe	1.00; U[0.40,0.70]		2.00; U[0.50,0.80]				2.00; U[0.20,0.40]		3.00; U[0.70,1.00]

Note:

1) In each shaded anchor element, the value refers to the mean value and, the distribution refers to the standard deviation distribution;
2) The n_c^3 states Low, Moderate and High are assigned with the values 1, 2 and 3, respectively.

Table 40 The anchor CPT of n_c^4 (System situation level)

Indication of condition		*i* = *j* (e.g., a)	*i* ≠ *j* (b)	*i* ≠ *j* (c)
HMI	Good	1.00; U[0.20,0.30]	2.00; U[0.40,0.60]	
	Normal			
	Misleading	2.00; U[0.40,0.60]	3.00; U[0.45,0.75]	

Note:

1) In each shaded anchor element, the value refers to the mean value and, the distribution refers to the standard deviation distribution;
2) The n_c^4 states Negligible, Moderate and Severe are assigned with the values 1, 2 and 3, respectively.

Table 41 The anchor CPT of n_c^5 (Diagnosis complexity)

Diagnosis procedure		Available	Normal	Incomplete
HMI	Good	1.00; U[0.20,0.30]		2.00; U[0.20,0.50]
	Normal			
	Misleading	1.00; U[0.30,0.60]		3.00; U[0.30,0.60]

Note:

1) In each shaded anchor element, the value refers to the mean value and, the distribution refers to the standard deviation distribution;
2) The n_c^5 states Obvious, Normal and Complex are assigned with the values 1, 2 and 3, respectively.

SECTION III: REFERENCES

This Section lists all the references cited in Sections I and II of this dissetation.

References

[1] Jazdi N. Cyber physical systems in the context of Industry 4.0. Automation, Quality and Testing, Robotics, 2014 IEEE International Conference on: IEEE; 2014. p. 1-4.

[2] Schwab K. The fourth industrial revolution: Crown Business; 2017.

[3] Zio E. The Future of Risk Assessment. Reliability Engineering & System Safety. 2018.

[4] Alur R. Principles of cyber-physical systems: MIT Press; 2015.

[5] Lee EA. Cyber Physical Systems: Design Challenges. IEEE Symposium on Object Oriented Real-Time Distributed Computing2008. p. 363-9.

[6] Colombo AW, Karnouskos S, Bangemann T. Towards the next generation of industrial cyber-physical systems. Industrial cloud-based cyber-physical systems: Springer; 2014. p. 1-22.

[7] Kim KD, Kumar PR. Cyber–Physical Systems: A Perspective at the Centennial. Proceedings of the IEEE. 2012;100:1287-308.

[8] Pajic M, Weimer J, Bezzo N, Sokolsky O, Pappas GJ, Lee I. Design and Implementation of Attack-Resilient Cyberphysical Systems: With a Focus on Attack-Resilient State Estimators. IEEE Control Systems. 2017;37:66-81.

[9] Ali S, Qaisar SB, Saeed H, Khan MF, Naeem M, Anpalagan A. Network challenges for cyber physical systems with tiny wireless devices: a case study on reliable pipeline condition monitoring. Sensors. 2015;15:7172-205.

[10] Liu K, Lee VCS, Ng KY, Chen J, Sang HS. Temporal Data Dissemination in Vehicular Cyber–Physical Systems. IEEE Transactions on Intelligent Transportation Systems. 2014;15:2419-31.

[11] Paelke V, Röcker C. User Interfaces for Cyber-Physical Systems: Challenges and Possible Approaches. International Conference of Design, User Experience, and Usability2015. p. 75-85.

[12] Wang W, Di Maio F, Zio E. Hybrid fuzzy-PID control of a nuclear Cyber-Physical System working under varying environmental conditions. Nuclear Engineering & Design. 2018;331:54-67.

[13] Lasi H, Fettke P, Kemper H-G, Feld T, Hoffmann M. Industry 4.0. Business & Information Systems Engineering. 2014;6:239-42.

[14] Bradley JM, Atkins EM. Optimization and Control of Cyber-Physical Vehicle Systems. Sensors. 2015;15:23020-49.

[15] Khaitan SK, McCalley JD. Design techniques and applications of cyberphysical systems: A survey. IEEE Systems Journal. 2015;9:350-65.

[16] Rajkumar RR, Lee I, Sha L, Stankovic J. Cyber-physical systems: the next computing revolution. Proceedings of the 47th design automation conference: ACM; 2010. p. 731-6.

[17] IAEA. Implementing Digital Instrumentation and Control Systems in the modernization of Nuclear Power Plants. Technical Report 2009;NP-T-1.4.

[18] Kriaa S, Pietre-Cambacedes L, Bouissou M, Halgand Y. A survey of approaches combining safety and security for industrial control systems. Reliability Engineering & System Safety. 2015;139:156-78.

[19] Piètre-Cambacédès L, Bouissou M. Cross-fertilization between safety and security engineering. Reliability Engineering & System Safety. 2013;110:110-26.

[20] Aven T. Identification of safety and security critical systems and activities. Reliability Engineering & System Safety. 2009;94:404-11.

[21] Zio E. Challenges in the vulnerability and risk analysis of critical infrastructures. Reliability Engineering & System Safety. 2016;152:137-50.

[22] Anderson PL, Geckil IK. Northeast blackout likely to reduce US earnings by $6.4 billion. Anderson Economic Group. 2003.

[23] Peng Z, Lu Y, Miller A, Johnson C, Zhao T. Risk assessment of railway transportation systems using timed fault trees. Quality and Reliability Engineering International. 2016;32:181-94.

[24] Singer P. Stuxnet and its hidden lessons on the ethics of cyberweapons. Case W Res J Int'l L. 2015;47:79.

[25] Varuttamaseni A, Bari R, Youngblood R. Construction of a Cyber Attack Model for Nuclear Power Plants. Brookhaven National Laboratory (BNL), Upton, NY (United States); 2017.

[26] Aven T. A unified framework for risk and vulnerability analysis covering both safety and security. Reliability engineering & System safety. 2007;92:745-54.

[27] Johnson C. CyberSafety: on the interactions between cybersecurity and the software engineering of safety-critical systems. Achieving System Safety. 2012:85-96.

[28] Zalewski J, Drager S, McKeever W, Kornecki AJ. Threat modeling for security assessment in cyberphysical systems. Proceedings of the Eighth Annual Cyber Security and Information Intelligence Research Workshop: ACM; 2013. p. 10.

[29] Miclea L, Sanislav T. About dependability in cyber-physical systems. Design & Test Symposium (EWDTS), 2011 9th East-West: IEEE; 2011. p. 17-21.

[30] Macana CA, Quijano N, Mojica-Nava E. A survey on cyber physical energy systems and their applications on smart grids. Innovative Smart Grid Technologies (ISGT Latin America), 2011 IEEE PES Conference on: IEEE; 2011. p. 1-7.

[31] Tidwell T, Gao X, Huang H-M, Lu C, Dyke S, Gill C. Towards configurable real-time hybrid structural testing: a cyber-physical system approach. Object/Component/Service-Oriented Real-Time Distributed Computing, 2009 ISORC'09 IEEE International Symposium on: IEEE; 2009. p. 37-44.

[32] Shin KG, Ramanathan P. Real-time computing: A new discipline of computer science and engineering. Proceedings of the IEEE. 1994;82:6-24.

[33] Kuo SM, Lee BH, Tian W. Real-time digital signal processing: fundamentals, implementations and applications: John Wiley & Sons; 2013.

[34] Chu C-T, Shih C-S. CPSSim: Simulation Framework for Large-Scale Cyber-Physical Systems. Cyber-Physical Systems, Networks, and Applications (CPSNA), 2013 IEEE 1st International Conference on: IEEE; 2013. p. 44-51.

[35] Lee EA. Cyber-physical systems-are computing foundations adequate. Position Paper for NSF Workshop On Cyber-Physical Systems: Research Motivation, Techniques and Roadmap: Citeseer; 2006. p. 1-9.

[36] Alur R. Principles of Cyber-Physical Systems. Mit Pr. 2015.

[37] Rajkumar R, Lee I, Sha L, Stankovic J. Cyber-physical systems: the next computing revolution. Design Automation Conference (DAC), 2010 47th ACM/IEEE: IEEE; 2010. p. 731-6.

[38] Rajkumar R, Lee I, Sha L, Stankovic J. Cyber-physical systems:the next computing revolution. Design Automation Conference2010. p. 731-6.

[39] Sztipanovits J, Koutsoukos X, Karsai G, Kottenstette N, Antsaklis P, Gupta V, et al. Toward a Science of Cyber–Physical System Integration. Proceedings of the IEEE. 2011;100:29-44.

[40] Sun B, Li X, Wan B, Wang C, Zhou X, Chen X. Definitions of predictability for cyber physical systems. Journal of Systems Architecture. 2016;63:48-60.

[41] Monostori L. Cyber-physical production systems: roots, expectations and R&D challenges. Procedia Cirp. 2014;17:9-13.

[42] Grund D, Reineke J, Wilhelm R. A template for predictability definitions with supporting evidence. OASIcs-OpenAccess Series in Informatics: Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik; 2011.

[43] Lee EA. Computing foundations and practice for cyber-physical systems: A preliminary report. University of California, Berkeley, Tech Rep UCB/EECS-2007-72. 2007.

[44] Zalewski J, Buckley IA, Czejdo B, Drager S, Kornecki AJ, Subramanian N. A Framework for Measuring Security as a System Property in Cyberphysical Systems. Information. 2016;7:33.

[45] Banerjee A, Venkatasubramanian KK, Mukherjee T, Gupta SKS. Ensuring safety, security, and sustainability of mission-critical cyber–physical systems. Proceedings of the IEEE. 2012;100:283-99.

[46] Wang EK, Ye Y, Xu X, Yiu S-M, Hui LCK, Chow K-P. Security issues and challenges for cyber physical system. Proceedings of the 2010 IEEE/ACM Int'l Conference on Green Computing and Communications & Int'l Conference on Cyber, Physical and Social Computing: IEEE Computer Society; 2010. p. 733-8.

[47] Mo Y, Kim TH-J, Brancik K, Dickinson D, Lee H, Perrig A, et al. Cyber–physical security of a smart grid infrastructure. Proceedings of the IEEE. 2012;100:195-209.

[48] Ashibani Y, Mahmoud QH. Cyber physical systems security: Analysis, challenges and solutions. Computers & Security. 2017;68:81-97.

[49] Wang W, Di Maio F, Zio E. Estimation of Failure on-Demand Probability and Malfunction Rate Values in Cyber-Physical Systems of Nuclear Power Plants. the 2017 International Topical Meeting on Probabilistic Safety Assessment and Analysis (PSA2017). Pittsburgh, USA2017.

[50] Rahman MS, Mahmud MA, Oo AMT, Pota HR. Multi-Agent Approach for Enhancing Security of Protection Schemes in Cyber-Physical Energy Systems. IEEE Transactions on Industrial Informatics. 2017;13:436-47.

[51] Li J, Huang X. Cyber Attack Detection of I&C Systems in NPPS Based on Physical Process Data. 2016 24th International Conference on Nuclear Engineering: American Society of Mechanical Engineers; 2016. p. V002T07A11-VT07A11.

[52] Kornecki AJ, Liu M. Fault Tree Analysis for Safety/Security Verification in Aviation Software. Electronics. 2013;2:41-56.

[53] Wang W, Di Maio F, Zio E. A Non-Parametric Cumulative Sum Approach for Online Diagnostics of Cyber Attacks to Nuclear Power Plants Resilience of Cyber-Physical Systems: From Risk Modelling to Threat Counteraction: Springer; 2018.

[54] Pedroni N, Zio E. An Adaptive Metamodel-Based Subset Importance Sampling approach for the assessment of the functional failure probability of a thermal-hydraulic passive system. Applied Mathematical Modelling. 2017;48:269-88.

[55] Pedroni N, Zio E, Apostolakis GE. Comparison of bootstrapped artificial neural networks and quadratic response surfaces for the estimation of the functional failure probability of a thermal–hydraulic passive system. Reliability Engineering & System Safety. 2010;95:386-95.

[56] Zio E, Pedroni N. Functional failure analysis of a thermal–hydraulic passive system by means of Line Sampling. Reliability Engineering & System Safety. 2009;94:1764-81.

[57] Eames DP, Moffett JD. The Integration of Safety and Security Requirements. International Conference on Computer Computer Safety, Reliability and Security1999. p. 468-80.

[58] Ropeik D, Gray GM. Risk: A practical guide for deciding what's really safe and what's dangerous in the world around you: Houghton Mifflin Harcourt; 2002.

[59] Boyes H. Trustworthy cyber-physical systems-a review. 2013.

[60] Anwar A, Mahmood AN. Cyber security of smart grid infrastructure. arXiv preprint arXiv:14013936. 2014.

[61] Wang W, Di Maio F, Zio E. Component- and system-level degradation modeling of digital Instrumentation and Control systems based on a Multi-State Physics Modeling Approach. Annals of Nuclear Energy. 2016;95:135-47.

[62] Zio E, Di Maio F. Processing dynamic scenarios from a reliability analysis of a nuclear power plant digital instrumentation and control system. Annals of Nuclear Energy. 2009;36:1386-99.

[63] Zaytoon J, Lafortune S. Overview of fault diagnosis methods for Discrete Event Systems. Annual Reviews in Control. 2013;37:308-20.

[64] Aldemir T, Guarro S, Mandelli D, Kirschenbaum J, Mangan LA, Bucci P, et al. Probabilistic risk assessment modeling of digital instrumentation and control systems using two dynamic methodologies. Reliability Engineering & System Safety. 2010;95:1011-39.

[65] Mcnelles P, Zeng ZC, Renganathan G, Lamarre G, Akl Y, Lu L. A comparison of Fault Trees and the Dynamic Flowgraph Methodology for the analysis of FPGA-based safety systems Part 1: Reactor trip logic loop reliability analysis. Reliability Engineering & System Safety. 2016;153:135-50.

[66] Machado RCS, Boccardo DR, Szwarcfiter JL. Software control and intellectual property protection in cyber-physical systems. Eurasip Journal on Information Security. 2016;2016:32.

[67] Jockenhövel-Barttfeld M, Taurines A, Hessler C. Quantification of Application Software Failures of Digital I&C in Probabilistic Safety Analyses. 13th International Conference on Probabilistic Safety Assessment and Management2016.

[68] Chen TM, Sanchez-Aarnoutse JC, Buford J. Petri net modeling of cyber-physical attacks on smart grid. IEEE Transactions on Smart Grid. 2011;2:741-9.

[69] Pasqualetti F, Dörfler F, Bullo F. Attack detection and identification in cyber-physical systems. IEEE Transactions on Automatic Control. 2013;58:2715-29.

[70] Shi D, Guo Z, Johansson KH, Shi L. Causality countermeasures for anomaly detection in cyber-physical systems. IEEE Transactions on Automatic Control. 2018;63:386-401.

[71] Yuan Y, Zhu Q, Sun F, Wang Q, Başar T. Resilient control of cyber-physical systems against Denial-of-Service attacks. International Symposium on Resilient Control Systems2013. p. 54-9.

[72] Zargar ST, Joshi J, Tipper D. A Survey of Defense Mechanisms Against Distributed Denial of Service (DDoS) Flooding Attacks. IEEE Communications Surveys & Tutorials. 2013;15:2046-69.

[73] Liang G, Zhao J, Luo F, Weller S, Dong ZY. A Review of False Data Injection Attacks Against Modern Power Systems. IEEE Transactions on Smart Grid. 2016;PP:1-.

[74] Mohammadpourfard M, Sami A, Seifi A. A Statistical Unsupervised Method Against False Data Injection Attacks: A Visualization-Based Approach. Expert Systems with Applications. 2017;84.

[75] Tan R, Nguyen HH, Foo EYS, Yau DKY, Kalbarczyk Z, Iyer RK, et al. Modeling and Mitigating Impact of False Data Injection Attacks on Automatic Generation Control. IEEE Transactions on Information Forensics & Security. 2017;12:1609-24.

[76] Trabelsi Z, Rahmani H. An Anti-Sniffer Based on ARP Cache Poisoning Attack. Information Systems Security. 2005;13:23-36.

[77] Ntalampiras S. Detection of Integrity Attacks in Cyber-Physical Critical Infrastructures Using Ensemble Modeling. IEEE Transactions on Industrial Informatics. 2015;11:104-11.

[78] Ntalampiras S. Automatic identification of integrity attacks in cyber-physical systems: Pergamon Press, Inc.; 2016.

[79] Shin J, Son H, Ur RK, Heo G. Development of a cyber security risk model using Bayesian networks. Reliability Engineering & System Safety. 2015;134:208-17.

[80] Xiang Y, Wang L, Liu N. Coordinated attacks on electric power systems in a cyber-physical environment. Electric Power Systems Research. 2017;149:156-68.

[81] Fang Y, Sansavini G. Optimizing Power System Investments and Resilience against Attacks. Reliability Engineering & System Safety. 2016;159:161-73.

[82] Hu X, Xu M, Xu S, Zhao P. Multiple Cyber Attacks Against a Target with Observation Errors and Dependent Outcomes: Characterization and Optimization. Reliability Engineering & System Safety. 2016;159:119-33.

[83] Yuan W, Zhao L, Zeng B. Optimal power grid protection through a defender–attacker–defender model. Reliability Engineering & System Safety. 2014;121:83-9.

[84] Wang W, Cammi A, Di Maio F, Lorenzi S, Zio E. A Monte Carlo-based exploration framework for identifying components vulnerable to cyber threats in nuclear power plants. Reliability Engineering & System Safety. 2018;175:24-37.

[85] Sha L, Gopalakrishnan S, Liu X, Wang Q. Cyber-physical systems: A new frontier. Sensor Networks, Ubiquitous and Trustworthy Computing, 2008 SUTC'08 IEEE International Conference on: IEEE; 2008. p. 1-9.

[86] Zio E. Prognostics and Health Management of Industrial Equipment. Volkswirtschaftliche Diskussionsbeiträge. 2013.

[87] Zio E. Integrated deterministic and probabilistic safety assessment: Concepts, challenges, research directions. Nuclear Engineering and Design. 2014;280:413-9.

[88] Aven T, Baraldi P, Flage R, Zio E. Uncertainty in risk assessment: the representation and treatment of uncertainties by probabilistic and non-probabilistic methods: John Wiley & Sons; 2013.

[89] Zio E, Pedroni N. Uncertainty characterization in risk analysis for decision-making practice: FonCSI; 2012.

[90] Paté - Cornell E. Finding and fixing systems weaknesses: Probabilistic methods and applications of engineering risk analysis. Risk Analysis. 2002;22:319-34.

[91] Paté - Cornell E. On "Black Swans" and "Perfect Storms": risk analysis and management when statistics are not enough. Risk Analysis: An International Journal. 2012;32:1823-33.

[92] Turati P, Pedroni N, Zio E. An adaptive simulation framework for the exploration of extreme and unexpected events in dynamic engineered systems. Risk analysis. 2017;37:147-59.

[93] Turati P, Pedroni N, Zio E. Simulation-based exploration of high-dimensional system models for identifying unexpected events. Reliability Engineering & System Safety. 2017;165:317-30.

[94] Zio E. An introduction to the basics of reliability and risk analysis: World scientific; 2007.

[95] McQueen MA, Boyer WF, Flynn MA, Beitel GA. Quantitative cyber risk reduction estimation methodology for a small SCADA control system. System Sciences, 2006 HICSS'06 Proceedings of the 39th Annual Hawaii International Conference on: IEEE; 2006. p. 226-.

[96] Sheyner O, Wing J. Tools for generating and analyzing attack graphs. International Symposium on Formal Methods for Components and Objects: Springer; 2003. p. 344-71.

[97] Ingols K, Lippmann R, Piwowarski K. Practical attack graph generation for network defense. Computer Security Applications Conference, 2006 ACSAC'06 22nd Annual: IEEE; 2006. p. 121-30.

[98] Fovino IN, Masera M, De Cian A. Integrating cyber attacks within fault trees. Reliability Engineering & System Safety. 2009;94:1394-402.

[99] Mitchell R, Chen R. Effect of intrusion detection and response on reliability of cyber physical systems. IEEE Transactions on Reliability. 2013;62:199-210.

[100] Backhaus S, Bent R, Bono J, Lee R, Tracey B, Wolpert D, et al. Cyber-physical security: A game theory model of humans interacting over control systems. IEEE Transactions on Smart Grid. 2013;4:2320-7.

[101] Xiang Y, Wang L, Zhang Y. Adequacy evaluation of electric power grids considering substation cyber vulnerabilities. International Journal of Electrical Power & Energy Systems. 2018;96:368-79.

[102] Zio E. Reliability engineering: Old problems and new challenges. Reliability Engineering & System Safety. 2009;94:125-41.

[103] Wang W, Di Maio F, Zio E. Three-Loop Monte Carlo Simulation Approach to Multi-State Physics Modeling for System Reliability Assessment. Reliability Engineering & System Safety. 2017;167.

[104] Banks DL, Aliaga JMR, Insua DR. Adversarial risk analysis: Chapman and Hall/CRC; 2015.

[105] Cox Jr LAT. Game theory and risk analysis. Risk Analysis. 2009;29:1062-8.

[106] Rios Insua D, Rios J, Banks D. Adversarial risk analysis. Journal of the American Statistical Association. 2009;104:841-54.

[107] Rothschild C, McLay L, Guikema S. Adversarial risk analysis with incomplete information: A level - k approach. Risk Analysis. 2012;32:1219-31.

[108] Moteff JD. Critical Infrastructure Resilience: The Evolution of Policy and Programs and Issues for Congress. Congressional Research Service Reports2012.

[109] Obama B. Presidential policy directive 21: Critical infrastructure security and resilience. Washington, DC. 2013.

[110] Polatidis N, Pavlidis M, Mouratidis H. Cyber-attack path discovery in a dynamic supply chain maritime risk management system. Computer Standards & Interfaces. 2018;56:74-82.

[111] Bi S, Zhang YJ. Graphical methods for defense against false-data injection attacks on power system state estimation. IEEE Transactions on Smart Grid. 2014;5:1216-27.

[112] Ge M, Hong JB, Yusuf SE, Kim DS. Proactive defense mechanisms for the software-defined Internet of Things with non-patchable vulnerabilities. Future Generation Computer Systems. 2018;78:568-82.

[113] Shandilya V, Simmons CB, Shiva S. Use of attack graphs in security systems. Journal of Computer Networks and Communications. 2014;2014.

[114] Kreps DM. Game theory and economic modelling: Oxford University Press; 1990.

[115] Nisan N, Roughgarden T, Tardos E, Vazirani VV. Algorithmic game theory: Cambridge University Press; 2007.

[116] Roger BM. Game theory: analysis of conflict. The President and Fellows of Harvard College, USA. 1991.

[117] Chen D, Xu M, Shi W. Defending a cyber system with early warning mechanism. Reliability Engineering & System Safety. 2018;169:224-34.

[118] Ma CY, Yau DK, Lou X, Rao NS. Markov game analysis for attack-defense of power networks under possible misinformation. IEEE Transactions on Power Systems. 2013;28:1676-86.

[119] Carl G, Kesidis G, Brooks RR, Rai S. Denial-of-Service Attack-Detection Techniques. IEEE Internet Computing. 2006;10:82-9.

[120] Debar H, Dacier M, Wespi A. Towards a taxonomy of intrusion-detection systems. Computer Networks. 1999;31:805-22.

[121] Mo Y, Chabukswar R, Sinopoli B. Detecting Integrity Attacks on SCADA Systems. Ifac Proceedings Volumes. 2011;44:11239-44.

[122] Tartakovsky AG, Rozovskii BL, Blažek RB, Kim H. Detection of intrusions in information systems by sequential change-point methods. Statistical Methodology. 2006;3:252-93.

[123] Teixeira A, Amin S, Sandberg H, Johansson KH. Cyber security analysis of state estimators in electric power systems. Decision and Control2010. p. 5991-8.

[124] Wald A. Sequential analysis: Courier Corporation; 1973.

[125] Hines JW, Garvey D. Development and Application of Fault Detectability Performance Metrics for Instrument Calibration Verification and Anomaly Detection. Journal of Pattern Recognition Research. 2006;1:2-15.

[126] Tartakovsky AG, Rozovskii BL, Blazek RB, Kim H. A novel approach to detection of intrusions in computer networks via adaptive sequential and batch-sequential change-point detection methods. IEEE Transactions on Signal Processing. 2006;54:3372-82.

[127] Page ES. Continuous inspection schemes. Biometrika. 1954;41:100-15.

[128] Viinikka J, Debar H, Mé L, Lehikoinen A, Tarvainen M. Processing intrusion detection alert aggregates with time series modeling. Information Fusion. 2009;10:312-24.

[129] Zhao X, Chu PS. Bayesian changepoint analysis for extreme events (typhoons, heavy rainfall, and heat waves): an RJMCMC approach. Journal of Climate. 2010;23:1034.

[130] Xie M, Goh TN, Ranjan P. Some effective control chart procedures for reliability monitoring. Reliability Engineering & System Safety. 2002;77:143-50.

[131] Maes MA, Fritzsons KE, Glowienka S. Structural robustness in the light of risk and consequence analysis. Structural engineering international. 2006;16:101-7.

[132] Beer T. Ecological risk assessment and quantitative consequence analysis. Human and Ecological Risk Assessment. 2006;12:51-65.

[133] Xiang Y, Wang L. A game-theoretic study of load redistribution attack and defense in power systems. Electric Power Systems Research. 2017;151:12-25.

[134] Wang W, Di Maio F, Zio E. A sensitivity analysis for the adequacy assessment of a multi-state physics modeling approach for reliability analysis. European Safety and Reliability Conference, Esrel2017. p. 465.

[135] Saltelli A, Chan K, Scott EM. Sensitivity analysis: Wiley New York; 2000.

[136] Di Maio F, Nicola G, Zio E, Yu Y. Ensemble-based sensitivity analysis of a Best Estimate Thermal Hydraulics model: Application to a Passive Containment Cooling System of an AP1000 Nuclear Power Plant. Annals of Nuclear Energy. 2014;73:200-10.

[137] Saltelli A, Ratto M, Andres T, Campolongo F, Cariboni J, Gatelli D, et al. Global Sensitivity Analysis. The Primer. 2008;304.

[138] Cadini F, Zio E, Di Maio F, Kopustinskas V, Urbonas R. A neural-network-based variance decomposition sensitivity analysis. International Journal of Nuclear Knowledge Management. 2007;2:299-312.

[139] Yu W, Harris T. Parameter uncertainty effects on variance-based sensitivity analysis. Reliability Engineering & System Safety. 2009;94:596-603.

[140] Sobol IM. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. Mathematics and computers in simulation. 2001;55:271-80.

[141] Borgonovo E. Measuring uncertainty importance: investigation and comparison of alternative approaches. Risk analysis. 2006;26:1349-61.

[142] Borgonovo E, Castaings W, Tarantola S. Model emulation and moment-independent sensitivity analysis: An application to environmental modelling. Environmental Modelling & Software. 2012;34:105-15.

[143] Wei P, Lu Z, Yuan X. Monte Carlo simulation for moment-independent sensitivity analysis. Reliability Engineering & System Safety. 2013;110:60-7.

[144] McKay M. Variance-based methods for assessing uncertainty importance in NUREG-1150 analyses. Los Alamos National Laboratory LA-UR-96-2695. 1996.

[145] Rocco CM, Zio E. Global sensitivity analysis in a multi-state physics model of component degradation based on a hybrid state-space enrichment and polynomial chaos expansion approach. IEEE Transactions on Reliability. 2013;62:781-8.

[146] Carlos S, Sánchez A, Ginestar D, Martorell S. Using finite mixture models in thermal-hydraulics system code uncertainty analysis. Nuclear Engineering and Design. 2013;262:306-18.

[147] Di Maio F, Nicola G, Zio E, Yu Y. Ensemble-based sensitivity analysis of a best estimate thermal hydraulics model: application to a passive containment cooling system of an AP1000 nuclear power plant. Annals of Nuclear Energy. 2014;73:200-10.

[148] Diaconis P, Zabell SL. Updating subjective probability. Journal of the American Statistical Association. 1982;77:822-30.

[149] Gibbs AL, Su FE. On choosing and bounding probability metrics. International statistical review. 2002;70:419-35.

[150] Zhang Y, Liu Y, Yang X. Parametric sensitivity analysis for importance measure on failure probability and its efficient Kriging solution. Mathematical Problems in Engineering. 2015;2015.

[151] Di Maio F, Colli D, Zio E, Tao L, Tong J. A Multi-State Physics Modeling approach for the reliability assessment of Nuclear Power Plants piping systems. Annals of Nuclear Energy. 2015;80:151-65.

[152] Li Y-F, Zio E, Lin Y-H. A multistate physics model of component degradation based on stochastic petri nets and simulation. IEEE Transactions on Reliability. 2012;61:921-31.

[153] Unwin SD, Lowry PP, Layton RF, Heasler PG, Toloczko MB. Multi-state physics models of aging passive components in probabilistic risk assessment. Pacific Northwest National Laboratory (PNNL), Richland, WA (US); 2011.

[154] Zio E. The Monte Carlo simulation method for system reliability and risk analysis: Springer; 2013.

[155] Zio E, Di Maio F, Tong J. Safety margins confidence estimation for a passive residual heat removal system. Reliability Engineering & System Safety. 2010;95:828-36.

[156] Di Maio F, Rai A, Zio E. A dynamic probabilistic safety margin characterization approach in support of Integrated Deterministic and Probabilistic Safety Analysis. Reliability Engineering & System Safety. 2016;145:9-18.

[157] Di Maio F, Picoco C, Zio E, Rychkov V. Safety margin sensitivity analysis for model selection in nuclear power plant probabilistic safety assessment. Reliability Engineering & System Safety. 2017;162:122-38.

[158] Wilks SS. Determination of sample sizes for setting tolerance limits. The Annals of Mathematical Statistics. 1941;12:91-6.

[159] Nutt WT, Wallis GB. Evaluation of nuclear safety from the outputs of computer codes in the presence of uncertainties. Reliability Engineering & System Safety. 2004;83:57-77.

[160] Wald A. An extension of Wilks' method for setting tolerance limits. The Annals of Mathematical Statistics. 1943;14:45-55.

[161] Wilks SS. Statistical prediction with special reference to the problem of tolerance limits. The annals of mathematical statistics. 1942;13:400-9.

[162] Beran R, Hall P. Interpolated nonparametric prediction intervals and confidence intervals. Journal of the Royal Statistical Society Series B (Methodological). 1993:643-52.

[163] Hutson AD. Calculating nonparametric confidence intervals for quantiles using fractional order statistics. Journal of Applied Statistics. 1999;26:343-53.

[164] Efron B, Tibshirani R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Statistical science. 1986:54-75.

[165] Sanchez-Saez F, Sánchez A, Villanueva J, Carlos S, Martorell S. Uncertainty analysis of a large break loss of coolant accident in a pressurized water reactor using non-parametric methods. Reliability Engineering & System Safety. 2018;174:19-28.

[166] Wang W, Di Maio F, Zio E. Adversarial Risk Analysis to Allocate Optimal Defense Resources for Protecting Nuclear Power Plants from Cyber Attacks. Risk Analysis. 2018.

[167] Wang W, Di Maio F, Zio E. Reliability Assessment of a Cyber Security Diagnostic Tool Embedded within a Nuclear Power Plant Considering the Uncertain Human Operator Cognition. 2018.

[168] Quijano EG, Insua DR, Cano J. Critical networked infrastructure protection from adversaries. Reliability Engineering & System Safety. 2016.

[169] Rios J, Insua DR. Adversarial risk analysis for counterterrorism modeling. Risk analysis. 2012;32:894-915.

[170] Li SY, Tang LC, Ng SH. Nonparametric CUSUM and EWMA Control Charts for Detecting Mean Shifts. Journal of Quality Technology. 2010;42:209-26.

[171] Yang SF, Cheng SW. A new non - parametric CUSUM mean chart. Quality & Reliability Engineering International. 2011;27:867-75.

[172] Alippi C, Boracchi G, Roveri M. Hierarchical change-detection tests. IEEE transactions on neural networks and learning systems. 2017;28:246-58.

[173] Qiu P, Hawkins D. A nonparametric multivariate cumulative sum procedure for detecting shifts in all directions. Journal of the Royal Statistical Society: Series D (The Statistician). 2003;52:151-64.

[174] Sorrells C, Qian L, Li H. Quickest detection of denial-of-service attacks in cognitive wireless networks. Homeland Security2012. p. 580-4.

[175] Kang J, Song YZ, Zhang JY. Accurate Detection of Peer-to-Peer Botnet using Multi-Stream Fused Scheme. Journal of Networks. 2011;6:807-14.

[176] Salem O, Vaton S, Gravey A. A scalable, efficient and informative approach for anomaly-based intrusion detection systems: theory and practice: John Wiley & Sons, Inc.; 2010.

[177] Baraldi P, Podofillini L, Mkrtchyan L, Zio E, Dang VN. Comparing the treatment of uncertainty in Bayesian networks and fuzzy expert systems used for a human reliability analysis application. Reliability Engineering & System Safety. 2015;138:176-93.

[178] Groth KM, Mosleh A. Deriving causal Bayesian networks from human reliability analysis data: A methodology and example model. Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability. 2012;226:361-79.

[179] Groth KM, Swiler LP. Bridging the gap between HRA research and HRA practice: A Bayesian network version of SPAR-H. Reliability Engineering & System Safety. 2013;115:33-42.

[180] Li PC, Zhang L, Dai LC, Li XF. Study on operator's SA reliability in digital NPPs. Part 3: A quantitative assessment method. Annals of Nuclear Energy. 2017;109:82-91.

[181] Mkrtchyan L, Podofillini L, Dang VN. Bayesian belief networks for human reliability analysis: A review of applications and gaps. Reliability Engineering & System Safety. 2015;139:1-16.

[182] Mkrtchyan L, Podofillini L, Dang VN. Methods for building Conditional Probability Tables of Bayesian Belief Networks from limited judgment: An evaluation for Human Reliability Application. Reliability Engineering & System Safety. 2016;151:93-112.

[183] Lee D-Y, Choi J-G, Lyou J. A safety assessment methodology for a digital reactor protection system. International Journal of Control, Automation, and Systems. 2006;4:105-12.

[184] Wang W, Zhao J, Tong J, Zhou J, Xiao P. Evaluation method of reliability indicator of reactor protection system. Atomic Energy Science and Technology. 2015;49:1101-8.

[185] Baraldi P, Di Maio F, Genini D, Zio E. Comparison of data-driven reconstruction methods for fault detection. IEEE Transactions on Reliability. 2015;64:852-60.

[186] Di Yun AMY, Vilim RB. Modeling the aging effects of Nuclear Power Plant resistance temperature detectors. 2012.

[187] Hashemian HM. Measurement of dynamic temperatures and pressures in nuclear power plants. 2011.

[188] Frogheri M, Alemberti A, Mansani L. The Lead Fast Reactor: Demonstrator (ALFRED) And ELFR Design. International Conference on FAST Reactors and Related Fuel Cycles: Safe Technologies and Sustainable Scenarios2013.

[189] Ponciroli R, Bigoni A, Cammi A, Lorenzi S, Luzzi L. Object-oriented modelling and simulation for the ALFRED dynamics. Progress in Nuclear Energy. 2014;71:15-29.

[190] Ponciroli R, Cammi A, Bona AD, Lorenzi S, Luzzi L. Development of the ALFRED reactor full power mode control system. Progress in Nuclear Energy. 2015;85:428-40.

[191] Skogestad S, Postlethwaite I. Multivariable feedback control: analysis and design: Wiley New York; 2007.

[192] EPRI. Advanced Light Water Reactor Utility Requirements Document, (Annex A reliability data base for passive ALWR PRAs). 2008.

[193] IAEA. Case study on the use of PSA methods: Assessment of technical specifications for the reactor protection system instrumentation. 1992.

[194] Kendall MG. The advanced theory of statistics. The advanced theory of statistics. 1946.

[195] Di Maio F, Compare M, Mattafirri S, Zio E. A double-loop Monte Carlo approach for Part Life Data Base reconstruction and scheduled maintenance improvement. Safety and Reliability: Methodology and Applications. 2014:1877-84.

[196] Ericson CA. Hazard analysis techniques for system safety: John Wiley & Sons; 2015.

[197] Spear R, Hornberger G. Eutrophication in Peel Inlet—II. Identification of critical uncertainties via generalized sensitivity analysis. Water Research. 1980;14:43-9.

[198] Borgonovo E, Hazen GB, Plischke E. A common rationale for global sensitivity measures and their estimation. Risk Analysis. 2016;36:1871-95.

[199] Kullback S, Leibler RA. On information and sufficiency. The annals of mathematical statistics. 1951;22:79-86.

[200] Balaban E, Saxena A, Bansal P, Goebel KF, Curran S. Modeling, detection, and disambiguation of sensor faults for aerospace applications. IEEE Sensors Journal. 2009;9:1907-17.

[201] Ding D, Wang Z, Han QL, Wei G. Security Control for Discrete-Time Stochastic Nonlinear Systems Subject to Deception Attacks. IEEE Transactions on Systems Man & Cybernetics Systems. 2018;48:779-89.

[202] Zhang H, Cheng P, Shi L, Chen J. Optimal DoS Attack Scheduling in Wireless Networked Control System. IEEE Transactions on Control Systems Technology. 2016;24:843-52.

[203] Boskvic JD, Mehra RK. Stable adaptive multiple model-based control design for accommodation of sensor failures. 2002;3:2046-51 vol.3.

[204] Di Maio F, Secchi P, Vantini S, Zio E. Fuzzy C-means clustering of signal functional principal components for post-processing dynamic scenarios of a nuclear power plant digital instrumentation and control system. IEEE Transactions on Reliability. 2011;60:415-25.

[205] Lehmann EL, Casella G. Theory of point estimation: Springer Science & Business Media; 2006.

[206] Mehetre DC, Roslin SE, Wagh SJ. Detection and prevention of black hole and selective forwarding attack in clustered WSN with Active Trust. Cluster Computing. 2018:1-16.

[207] Noureddine MA, Marturano A, Keefe K, Bashir M, Sanders WH. Accounting for the Human User in Predictive Security Models. Dependable Computing (PRDC), 2017 IEEE 22nd Pacific Rim International Symposium on: IEEE; 2017. p. 329-38.

[208] Institute P. Cost of cyber crime study: insights on the security investments that make a difference. 2017.

[209] Viscusi WK. Valuing risks of death from terrorism and natural disasters. Journal of Risk and Uncertainty. 2009;38:191-213.

[210] Viscusi WK, Aldy JE. The value of a statistical life: a critical review of market estimates throughout the world. Journal of risk and uncertainty. 2003;27:5-76.

[211] Nazir S, Patel S, Patel D. Assessing and augmenting SCADA cyber security: A survey of techniques. Computers & Security. 2017;70:436-54.

[212] Nespoli P, Papamartzivanos D, Mármol FG, Kambourakis G. Optimal countermeasures selection against cyber attacks: A comprehensive survey on reaction frameworks. IEEE Communications Surveys & Tutorials. 2017.

[213] Yang Q, Yang J, Yu W, An D, Zhang N, Zhao W. On false data-injection attacks against power system state estimation: Modeling and countermeasures. IEEE Transactions on Parallel and Distributed Systems. 2014;25:717-29.

[214] De Roze BC, Nyman TH. The software life cycle—A management and technological challenge in the Department of Defense. IEEE Transactions on Software Engineering. 1978:309-18.

[215] Bier V, Oliveros S, Samuelson L. Choosing what to protect: Strategic defensive allocation against an unknown attacker. Journal of Public Economic Theory. 2007;9:563-87.

[216] Fang Y, Sansavini G. Optimizing power system investments and resilience against attacks. Reliability Engineering & System Safety. 2017;159:161-73.

[217] Levitin G. Optimal defense strategy against intentional attacks. IEEE Transactions on Reliability. 2007;56:148-57.

[218] Levitin G, Hausken K. Parallel systems under two sequential attacks. Reliability Engineering & System Safety. 2009;94:763-72.

[219] Zhang J, Zhuang J, Jose VRR. The role of risk preferences in a multi-target defender-attacker resource allocation game. Reliability Engineering & System Safety. 2018;169:95-104.

[220] Osborne MJ, Rubinstein A. A course in game theory: MIT press; 1994.

[221] Zhuang J, Bier VM. Balancing terrorism and natural disasters—Defensive strategy with endogenous attacker effort. Operations Research. 2007;55:976-91.

[222] Zhuang J, Bier VM. Secrecy and deception at equilibrium, with applications to anti‐terrorism resource allocation. Defence and Peace Economics. 2011;22:43-61.

[223] Wurm J, Jin Y, Liu Y, Hu S, Heffner K, Rahman F, et al. Introduction to cyber-physical system security: A cross-layer perspective. IEEE Trans Multi-Scale Comput Syst. 2017;3:215-27.

[224] Zou L-L. Risk Analysis of Cyber Security in Nuclear Power Plant. Nuclear Power Plants: Innovative Technologies for Instrumentation and Control Systems: The Second International Symposium on Software

Reliability, Industrial Safety, Cyber Security and Physical Protection of Nuclear Power Plant: Springer; 2017. p. 139.

[225] Cano J, Insua DR, Tedeschi A, Turhan Uu. Security economics: an adversarial risk analysis approach to airport protection. Annals of Operations Research. 2016;245:359-78.

[226] Modarres M. Risk analysis in engineering: techniques, tools, and trends: CRC press; 2016.

[227] Bernoulli D. Exposition of a new theory on the measurement of risk. The Kelly Capital Growth Investment Criterion: Theory and Practice: World Scientific; 2011. p. 11-24.

[228] Von Neumann J, Morgenstern O. Theory of games and economic behavior (commemorative edition): Princeton university press; 2007.

[229] Bricha N, Nourelfath M. Critical supply network protection against intentional attacks: A game-theoretical model. Reliability Engineering & System Safety. 2013;119:1-10.

[230] Grechuk B, Zabarankin M. Inverse portfolio problem with coherent risk measures. European Journal of Operational Research. 2016;249:740-50.

[231] Hershey JC, Schoemaker PJ. Probability versus certainty equivalence methods in utility measurement: Are they equivalent? Management Science. 1985;31:1213-31.

[232] Cox J, Sadiraj V. Small-and large-stakes risk aversion: Implications of concavity calibration for decision theory. 2005.

[233] Pratt JW. Risk aversion in the small and in the large. HANDBOOK OF THE FUNDAMENTALS OF FINANCIAL DECISION MAKING: Part I: World Scientific; 2013. p. 317-31.

[234] Maskin E. The theory of implementation in Nash equilibrium: A survey. Social goals and social organization. 1985:173-204.

[235] Gibbons R. A primer in game theory: Harvester Wheatsheaf; 1992.

[236] Nash J. Non-cooperative games. Annals of mathematics. 1951:286-95.

[237] Walker JJ. Cyber security concerns for emergency management. Emergency Management: InTech; 2012.

[238] Authen S, Holmberg J-E. Reliability analysis of digital systems in a probabilistic risk analysis for nuclear power plants. Nuclear Engineering and Technology. 2012;44:471-82.

[239] Gray RM, Neuhoff DL. Quantization. IEEE Transactions on Information Theory. 1998;44:2325-83.

[240] Widrow B. Statistical analysis of amplitude-quantized sampled-data systems. Transactions of the American Institute of Electrical Engineers Part II Applications & Industry. 2012;79:555-68.

[241] Jones HL. Failure detection in linear systems: Massachusetts Institute of Technology; 1973.

[242] Tian E, Yue D. Reliable H∞ filter design for T‐S fuzzy model‐based networked control systems with random sensor failure. International Journal of Robust and Nonlinear Control. 2013;23:15-32.

[243] Duda RO, Hart PE. Pattern Classification, Scene Analysis. 1973.

[244] Di Maio F, Baraldi P, Zio E, Seraoui R. Fault Detection in Nuclear Power Plants Components by a Combination of Statistical Methods. IEEE Transactions on Reliability. 2013;62:833-45.

[245] Al-Dahidi S, Baraldi P, Di Maio F, Zio E. A novel fault detection system taking into account uncertainties in the reconstructed signals. Annals of Nuclear Energy. 2014;73:131-44.

[246] Jardine AKS, Lin D, Banjevic D. A review on machinery diagnostics and prognostics implementing condition-based maintenance. Mechanical Systems & Signal Processing. 2006;20:1483-510.

[247] Lee SJ, Man CK, Seong PH. An analytical approach to quantitative effect estimation of operation advisory system based on human cognitive process using the Bayesian belief network. Reliability Engineering & System Safety. 2008;93:567-77.

[248] Nazir S, Patel S, Patel D. Assessing and Augmenting SCADA Cyber Security-A Survey of Techniques. Computers & Security. 2017;70.

[249] Gratian M, Bandi S, Cukier M, Dykstra J, Ginther A. Correlating Human Traits and Cybersecurity Behavior Intentions. Computers & Security. 2017;73:345–58.

[250] Kim HE, Han SS, Kim J, Kang HG. Systematic development of scenarios caused by cyber-attack-induced human errors in nuclear power plants. Reliability Engineering & System Safety. 2017;167:290-301.

[251] Commission USNR. Technical Basis and Implementation Guidelines for A Technique for Human Event Analysis (ATHEANA) Washington, DC 20555-0001 2000.

[252] Groth KM, Mosleh A. Deriving causal Bayesian networks from human reliability analysis data: A methodology and example model. Proceedings of the Institution of Mechanical Engineers Part O Journal of Risk & Reliability. 2012;226:361-79.

[253] Zou Y, Zhang L, Li P. Reliability forecasting for operators' situation assessment in digital nuclear power plant main control room based on dynamic network model. Safety science. 2015;80:163-9.

[254] Commission UNR. Technical basis and implementation guidelines for a technique for human event analysis (ATHEANA). NUREG-1624, Rev. 2000;1.

[255] Kim MC, Seong PH. A method for identifying instrument faults in nuclear power plants possibly leading to wrong situation assessment. Reliability Engineering & System Safety. 2008;93:316-24.

[256] Naderpour M, Lu J, Zhang G. A human-system interface risk assessment method based on mental models. Safety science. 2015;79:286-97.

[257] Kim AR, Kim JH, Jang I, Seong PH. A framework to estimate probability of diagnosis error in NPP advanced MCR. Annals of Nuclear Energy. 2018;111:31-40.

[258] John M. O'Hara WSB, Paul M. Lewis and J.J. Persensky. The Effects of Interface Management Tasks On Crew Performance and Safety in Complex, Computer-Based Systems. Washington, DC 20555-0001: U.S. Nuclear Regulatory Commission; 2002.

[259] Kim Y, Park J, Jung W. A quantitative measure of fitness for duty and work processes for human reliability analysis. Reliability Engineering & System Safety. 2017;167:595-601.

[260] Park J, Jung J-Y, Jung W. The use of a process mining technique to characterize the work process of main control room crews: A feasibility study. Reliability Engineering & System Safety. 2016;154:31-41.

[261] Gertman D, Blackman H, Marble J, Byers J, Smith C. The SPAR-H human reliability analysis method. US Nuclear Regulatory Commission. 2005.

[262] Kim Y, Park J, Jung W, Choi SY, Kim S. Estimating the Quantitative Relation between PSFs and HEPs from Full-Scope Simulator Data. Reliability Engineering & System Safety. 2018.

[263] John Forester HL, Vinh N. Dang, Andreas Bye, Erasmia Lois, Mary Presley, Julie Marble, Rod Nowell, Helena Broberg, Michael Hildenbrandt, Bruce Hallbert, and Tommy Morgan. The U.S. HRA Empirical Study – Assessment of HRA Method Predictions against Operating Crew Performance on a U.S. Nuclear Power Plant Simulator. Washington, DC 20555-0001: U.S. Nuclear Regulatory Commission; 2016.

[264] Hallbert B. The employment of empirical data and Bayesian methods in human reliability analysis: a feasibility study: US Nuclear Regulatory Commission, Office of Nuclear Regulatory Research; 2007.

SECTION IV: PUBLICATIONS

This Section corrects the papers related to this Ph.D. research activity. The research activity has led to the acceptance (5), the submission (1) and the work in progress (1) of 7 manuscripts at international peer-reviewed journals and books (see Table 42), and the acceptance of 4 papers presented at the international academic conferences (see Table 43).

One conference paper, five journal papers and one book chapter (as mentioned in Section 1.6) which introduce the core techniques of this Ph.D. work are attached in this Section.

Table 42 Publications in international peer-reviewed journals and books

No.	Title	Authors	Status
1[B]	A Non-Parametric Cumulative Sum Approach for Real-Time Diagnostics of Cyber Attacks to Nuclear Power Plants	W. Wang, F. Di Maio, E. Zio	Book chapter of "*Resilience of Cyber-Physical Systems: From Risk Modeling to Threat Counteraction*", DOI: 10.1007/978-3-319-95597-1
1[J]	Component- and System-Level Degradation Modeling of Digital Instrumentation and Control Systems Based on a Multi-State Physics Modeling Approach	W. Wang, F. Di Maio, E. Zio	*Annals of Nuclear Energy 95 (2016) 135–147*
2[J]	Three-Loop Monte Carlo Simulation Approach to Multi-State Physics Modeling for System Reliability Assessment	W. Wang, F. Di Maio, E. Zio	*Reliability Engineering and System Safety 167 (2017) 276–289*
3[J]	A Monte Carlo-based Exploration Framework for Identifying Components Vulnerable to Cyber Threats in Nuclear Power Plants	W. Wang, A. Cammi, F. Di Maio, S. Lorenzi, E. Zio	*Reliability Engineering and System Safety 175 (2018) 24–37*
4[J]	Adversarial Risk Analysis to Allocate Optimal Defense Resources for Protecting Nuclear Power Plants from Cyber Attacks	W. Wang, F. Di Maio, E. Zio	*Risk Analysis, under review*
5[J]	Reliability Assessment of an Online Cyber Security Diagnostic Tool of a Nuclear Power Plant under Uncertain Human Operator Cognition	W. Wang, F. Di Maio, E. Zio	Work in progress
6[J]	A Hybrid Fuzzy-PID Controller for an Intelligent Cyber-Physical System Under Varying Environmental Conditions	W. Wang, F. Di Maio, E. Zio	*Nuclear Engineering and Design 331 (2018) 54-67*
7[J]	Analysis of Data Errors in Communication-based Train Control Systems	T. Wang, W. Wang, E. Zio, T. Tang, D. Zhou	Work in progress

Table 43 Publications in international conferences

No.	Title	Authors	Status
1[C]	A Sensitivity Analysis for the Adequacy Assessment of a Multi-State Physics Modeling Approach for Reliability Analysis	W. Wang, F. Di Maio, E. Zio	The 2016 European Safety and Reliability Conference (ESREL2016)
2[C]	Estimation of Failure on-Demand Probability and Malfunction Rate Values in Cyber-Physical Systems of Nuclear Power Plants	W. Wang, F. Di Maio, E. Zio	The 2017 International Topical Meeting on Probabilistic Safety Assessment and Analysis (PSA2017)
3[C]	Integrating Physical Degradation Modeling within the Seismic Fragility Analysis of Nuclear Power Plant Equipment	W. Wang, S. Zhang	The 2017 International Topical Meeting on Probabilistic Safety Assessment and Analysis (PSA2017)
4[C]	A Hybrid Monte Carlo and Possibilistic Approach to Estimate Non-Suppression Probability in Fire Probabilistic Safety Analysis	W. Wang, Di Maio, F., Baraldi, P., Zio, E.	The 2017 2nd International Conference on System Reliability and Safety (ICSRS2017)

PAPER 1[C]: A SENSITIVITY ANALYSIS FOR THE ADEQUACY ASSESSMENT OF A MULTI-STATE PHYSICS MODELING APPROACH FOR RELIABILITY ANALYSIS

Wei Wang, Francesco Di Maio, Enrico Zio

In the 26th European Safety and Reliability Conference (ESREL 2016)
Glasgow, United Kingdom, 25-29 September 2016
Page 465

CRC Press
Taylor & Francis Group

Resilience Risk, Reliability and Safety: Innovating Theory and Practice - Proceedings of the 26th European Safety and Reliability Conference, ESREL 2016

Editors: Lesley Walls, Matthew Revie, Tim Bedford

ISBN: 9781138029972 - CAT# K30324

PAPER 1[J]: COMPONENT- AND SYSTEM-LEVEL DEGRADATION MODELING OF DIGITAL INSTRUMENTATION AND CONTROL SYSTEMS BASED ON A MULTI-STATE PHYSICS MODELING APPROACH

Wei Wang, Francesco Di Maio*, Enrico Zio

Annals of Nuclear Energy 95 (2016) 135–147

https://doi.org/10.1016/j.anucene.2016.05.006

PAPER 2[J]: THREE-LOOP MONTE CARLO SIMULATION APPROACH TO MULTI-STATE PHYSICS MODELING FOR SYSTEM RELIABILITY ASSESSMENT

Wei Wang, Francesco Di Maio*, Enrico Zio

Reliability Engineering and System Safety 167 (2017) 276–289

https://doi.org/10.1016/j.ress.2017.06.003

PAPER 3[J]: A MONTE CARLO-BASED EXPLORATION FRAMEWORK FOR IDENTIFYING COMPONENTS VULNERABLE TO CYBER THREATS IN NUCLEAR POWER PLANTS

Wei Wang, Antonio Cammi, Francesco Di Maio*, Stefano Lorenzi, Enrico Zio

Reliability Engineering and System Safety 175 (2018) 24–37

https://doi.org/10.1016/j.ress.2018.03.005

PAPER 4[J]: ADVERSARIAL RISK ANALYSIS TO ALLOCATE OPTIMAL DEFENSE RESOURCES FOR PROTECTING NUCLEAR POWER PLANTS FROM CYBER ATTACKS

Wei Wang, Francesco Di Maio*, Enrico Zio

Risk Analysis, *under review*

PAPER 1[B]: A NON-PARAMETRIC CUMULATIVE SUM APPROACH FOR REAL-TIME DIAGNOSTICS OF CYBER ATTACKS TO NUCLEAR POWER PLANTS

Wei Wang, Francesco Di Maio*, Enrico Zio

[Chapter 9]

Springer

© 2018 **Resilience of Cyber-Physical Systems: From Risk Modelling to Threat Counteraction**

Series Title: Advanced Sciences and Technologies for Security Applications

Editors: Francesco Flammini

DOI: 10.1007/978-3-319-95597-1

PAPER 5[J]: RELIABILITY ASSESSMENT OF AN ONLINE CYBER SECURITY DIAGNOSTIC TOOL OF A NUCLEAR POWER PLANT UNDER UNCERTAIN HUMAN OPERATOR COGNITION

Wei Wang, Francesco Di Maio*, Enrico Zio

Work in progress

[This page intentionally left blank.]