Considering the Human Operator Cognitive Process for the Interpretation of Diagnostic Outcomes Related to Component Failures and Cyber Security Attacks

Wei Wang[1], Francesco Di Maio[2], Enrico Zio[2,3,4]

[1]*Department of Mechanical Engineering, City University of Hong Kong, Kowloon, Hong Kong, China*

[2]*Department of Energy, Politecnico di Milano, Via La Masa 34, 20156 Milano, Italy*

[3]*MINES ParisTech / PSL Université Paris, Centre de Recherche sur les Risques et les Crises (CRC), Sophia Antipolis, France*

[4]*Eminent Scholar, Department of Nuclear Engineering, Kyung Hee University*

Abstract: In this work, we consider diagnostics of cyber attacks in Cyber-Physical Systems (CPSs), based on data analytics. For the first time to authors knowledge, the performance of such diagnosis is quantified considering the possible failure of the human operator cognitive process in interpreting and understanding the diagnosis support tool outcomes.

A Non-Parametric CUmulative SUM (NP-CUSUM) approach is used for data-driven diagnostic, and the cognitive process of the human operator who interprets its outputs is modelled by a Bayesian Belief Network (BBN). The overall framework is applied on the digital controller of the Advanced Lead-cooled Fast Reactor European Demonstrator (ALFRED).

Keywords: Cyber-Physical System; Diagnostic; Non-Parametric CUmulative SUM (NP-CUSUM); Human Cognition; Bayesian Belief Network; Nuclear Power Plant.

ABBREVIATIONS

ALFRED	Advanced Lead-cooled Fast Reactor European Demonstrator
BBN	Bayesian Belief Network
CPD	Conditional Probability Distribution
CPS	Cyber-Physical System
CPT	Conditional Probability Table
CR	Control Rod
DAC	Digital-to-Analog Converter
DoS	Denial of Service
HMI	Human-Machine Interface
I&C	Instrumentation and Control
LSB	Least Significant Bit
MC	Monte Carlo
NP-CUSUM	Non-Parametric CUmulative SUM
NPP	Nuclear Power Plant
PI	Proportional-Integral
PSF	Performance Shaping Factor
SG	Steam Generator
SISO	Single Input Single Output

NOMENCLATURE

P_{Th}	Thermal power
h_{CR}	Height of control rods
$T_{L,hot}$	Coolant core outlet temperature
$T_{L,cold}$	Coolant SG outlet temperature
Γ	Coolant mass flow rate
T_{feed}	Feedwater SG inlet temperature
T_{steam}	Steam SG outlet temperature
p_{SG}	SG pressure
G_{water}	Feedwater mass flow rate
G_{att}	Attemperator mass flow rate
kv	Turbine admission valve coefficient
P_{Mech}	Mechanical power
t	Time
t_R	Accident time
t_M	Mission time

dt	Sensor measuring time interval
y	Variable (safety parameter)
y^{ref}	Reference value of controller set point value of y
$y^{sensor}(t)$	Sensor measurement
$y^{F,sensor}(t)$	Sensor false measurement
$y^{feed}(t)$	Measurement received by the computing (feeding) subsystem
$y^{monitor}(t)$	Measurement received by the monitoring subsystem
$\delta_y(t)$	Sensor measuring error
$q_y(t)$	Converter quantization error
a	Accidental scenario
b	Bias factor
$Y(t)$	Redundant channel measure, $Y = y^{feed}$ and $y^{monitor}$
$S_Y(t)$	Score function-based statistic of the collected $Y(t)$, $S_Y(t) = S_y^{feed}(t)$ and $S_y^{monitor}(t)$
h_y	Positive threshold
τ_Y	Time to alarm, $\tau_Y = \tau_y^{feed}$ and $\tau_y^{monitor}$
$\Delta\tau_Y$	Delay difference between τ_y^{feed} and $\tau_y^{monitor}$
Γ_y^{ref}	Reference delay difference
ε_y	NP-CUSUM tuning parameter
n_p^{α}	α-th parent node, α=1, 2, …, 7
$S_p^{\alpha,\gamma}$	γ-th state of α-th parent node, γ=1, 2 or 3
$p(S_p^{\alpha,\gamma})$	Probability of the occurrence of $S_p^{\alpha,\gamma}$
n_c^{β}	β-th child node, β=1, 2, …, 5
$S_c^{\beta,\gamma}$	γ-th state of β-th child node, γ=1, 2 or 3
$p(S_c^{\beta,\gamma})$	Probability of the occurrence of $S_c^{\beta,\gamma}$
i	NP-CUSUM online assignment
j	Real accidental event
k	Operator diagnostic cognitive decision
$p(j\|i)$	Probability of occurrence of an accidental event j, conditional on its online assignment i
$p(k\|j,i)$	Conditional probability of k, conditional on the combination (j,i)
$p(j,k\|i)$	Conditional probability of the operator diagnostic cognitive decision is k whose real event is j, when interpreting the online outcome i
p^i_{correct}	Probability of correct diagnostic conditional on the online assignment i
$e_c^{\beta,\theta}$	Elements at the anchor CPT of n_c^{β}, θ=anchor for selected anchors

with empirical distributions $N\left(u\left(e_c^{\beta,\theta=\text{anchor}}\right),\sigma\left(e_c^{\beta,\theta=\text{anchor}}\right)\right)$, and θ=other for the other elements whose missing mean and standard deviation values shall be linearly interpolated

1. INTRODUCTION

Cyber-Physical Systems (CPSs) feature a tight combination of (and coordination between) the physical process that runs in the system and the cyber domain that, by high automation level, real-time monitors, dynamically controls and supports decision-making during system operations [1-3].

Despite the benefits of CPSs, such as increased functionality, expanded capability and improved flexibility, the concern exists that their operation can be compromised not only by failures but also by attacks [4-6], that can be both physical or cyber.

Attacks depend on many factors (e.g., attacker profile, skills, motivation, etc.), which makes it difficult for defenders to anticipate and diagnose attack scenarios [4, 7, 8]. This is particularly true for cyber attacks, which are the focus of this work, since the majority of game-theoretic models assume that the defender moves first (e.g. designing a system, as in this work), and that the attacker moves after [9-13]. However, this means that an attacker can maximize the objective (of his/her malevolent act) and cyber attacks might be disguised from random failures, rendering the recovery difficult [14, 15].

Data-driven methods (e.g., the Sequential Probability Ratio Test (SPRT) [16, 17], the Cumulative Sum (CUSUM) chart [14, 18, 19], the Exponentially Weighted Moving Average (EWMA) inspection scheme [20]) have been proposed for the analysis of deviations in the observations from nominal values for diagnosing component stochastic failures. Machine learning techniques, including supervised learning (e.g., Support Vector Machine (SVM) [21], Neural Network (NN) [22]), unsupervised learning [23] and reinforcement learning (e.g., Q-learning [24, 25]), have also been proposed for such diagnosing task [26, 27].

Practically, the outcomes of the diagnosis are made available to operators via digital Human-Machine Interfaces (HMIs). The operator is requested to interpret these outcomes and take decisions on what to do or not do for responding to the effects induced by the diagnosed failures [28-30]. The human cognition process for assessing the system state based on the interpretation of the diagnostic outcomes can improve the diagnostic performance or worsen it [28, 29, 31-34]. This has been analyzed considered

in the literature, using expert judgment [35, 36] and artificial intelligence [37, 38].

Methods and algorithms have been developed also for diagnosing cyber attacks [14, 39, 40]. In this work, without loss of generality, a Non-Parametric CUmulative SUM (NP-CUSUM) method (a sequential anomaly detection technique proposed in the literature for detecting parameter changes in physical systems [14, 41, 42]) is adopted for components failures [43, 44] and cyber attacks [14, 45-47], and the human operator cognition process that interprets the monitoring/detection outcomes for situation assessment (i.e., the operator develops his/her mental representation of the specific current situation), and response planning (i.e., the operator takes decisions for dealing with the assessed situation) [29, 30, 48] is originally modelled by a Bayesian Belief Network (BBN). Specifically, a BBN typically used for structuring expert knowledge, understanding and cognition errors related to component stochastic failures diagnosis [28, 30, 31, 49-53] is here originally tailored for modelling the human operator cognitive process for interpreting of the diagnostic outcomes originated from cyber attacks.

A further novelty of the work consists in the overall framework of analysis, shown in Fig. 1, structured to capitalize the information made available by monitoring a CPS affected by cyber attacks and/or component stochastic failures for diagnosing the occurring events by a data-driven diagnostic tool, such as NP-CUSUM, where considering the human operator cognitive process in the interpretation of the diagnostic outcomes that influence the operator decision.

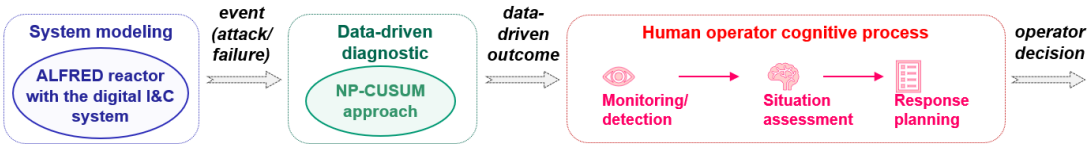

Fig. 1 Overall framework

A case study is considered, concerning stochastic components failures and cyber attacks that can occur in the digital Instrumentation and Control (I&C) system of the

Advanced Lead Fast Reactor European Demonstrator (ALFRED) [54]. An object-oriented simulator previously developed [55, 56], comprising a multi-loop Proportional-Integral (PI) controller [57], is utilized for simulating the ALFRED dynamic response to failures and cyber attacks. Data are fed to the NP-CUSUM algorithm [15], and the diagnostic outcomes are interpreted by operators, whose cognitive process is modelled by BBN.

The rest of the paper is organized as follows. Section 2 presents the main characteristics of the ALFRED reactor with its digital I&C system, the MC engine for injection of components failures and cyber breaches, and the NP-CUSUM technique. The operator cognitive process modelled by BBN is presented in Section 3. Section 4 presents the results and Section 5 concludes the paper.

2. THE ADVANCED LEAD-COOLED FAST REACTOR EUROPEAN DEMONSTRATOR

2.1 The Reactor and the digital I&C system

ALFRED is a small-size (300 MW) pool-type fast reactor, cooled by molten lead [54]. During operation, Control Rods (CRs) height h_{CR} is adjusted for thermal power (P_{Th}) regulation, reactivity swing compensation during the cycle, and scram for safe shutdown when necessary [58].

At full power nominal conditions, the dynamics processing of the primary and secondary cooling systems is controlled by a multi-loop PI (Proportional and Integral) control scheme (see Fig. 2). Such decentralized control scheme allows simplicity of implementation and robustness to malfunctioning of the single control loops [55, 56]. Both feedback and feedforward digital control schemes are used (see Fig. 2 shadowed part). The PI-based feedback control configuration employs four SISO (Single Input Single Output) control loops independent of each other.

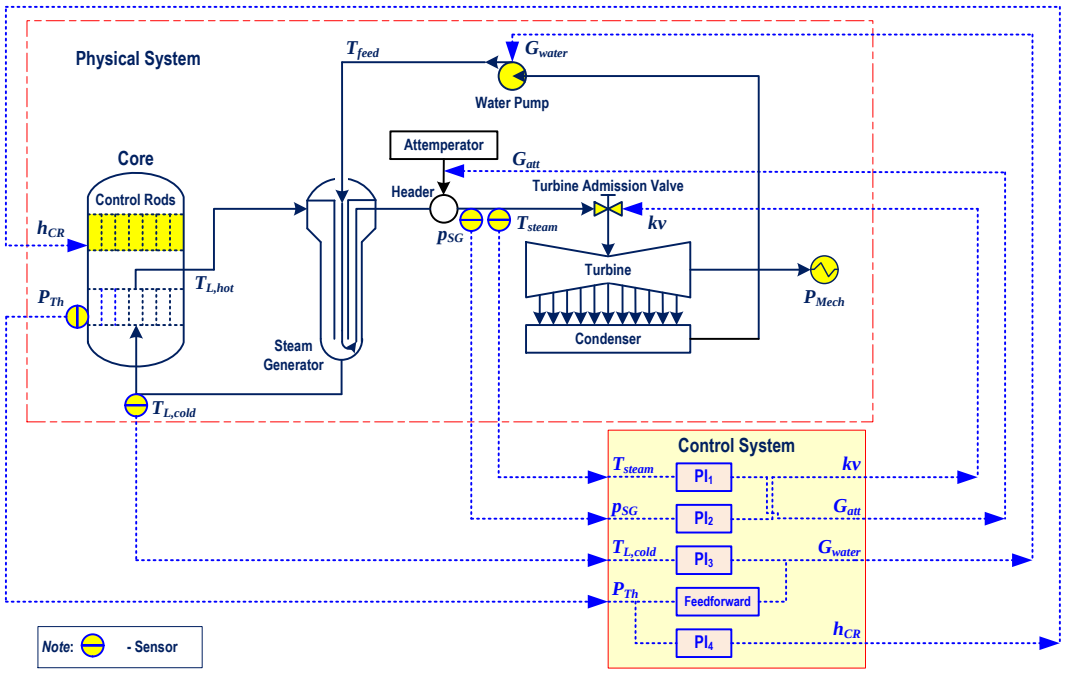

Fig. 2. ALFRED reactor control scheme

The control system aims at keeping the controlled variables at the steady state values, which give the optimal working conditions at full power nominal conditions. For example, it is expected that the coolant (i.e., lead) flow coming from the cold pool enters the core at temperature $T_{L,cold}$ equal to 400 °C, controlled by the $T_{L,cold}$-PI$_3$-G_{water} control loop of Fig. 2.

The parameters specification at full power nominal conditions are reported in Table 1.

Table 1 ALFRED parameters values at full power nominal conditions

Parameter	Parameter Description	Value	Unit
P_{Th}	Thermal power	$300 \cdot 10^6$	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	$kg \cdot s^{-1}$
T_{feed}	Feedwater SG inlet temperature	335	°C
T_{steam}	Steam SG outlet temperature	450	°C
p_{SG}	SG pressure	$180 \cdot 10^5$	Pa
G_{water}	Feedwater mass flow rate	192	$kg \cdot s^{-1}$
G_{att}	Attemperator mass flow rate	0.5	$kg \cdot s^{-1}$
kv	Turbine admission valve coefficient	1	-
P_{Mech}	Mechanical power	$146 \cdot 10^6$	W

Redundancy is commonly applied to sensors and signal processing units of a

digital I&C system [59]. In the ALFRED digital control scheme, redundancy has been used to design each independent SISO loop.

Fig. 3 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 $T_{L,cold}^{feed}(t)$ are 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.

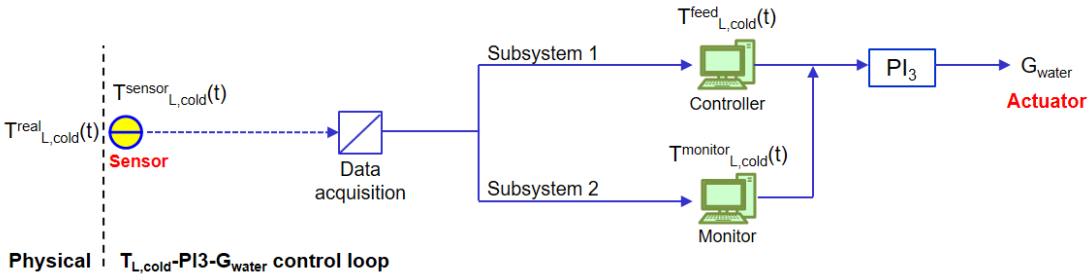

Fig. 3 The redundancy design of the $T_{L,cold}$-PI_3-G_{water} control loop

Measurements are realistically considered to be affected by two types of errors [60, 61]: 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 2 lists the reference values of the controlled variables, the distributions of sensor measurement errors and the quantization errors that each control loop is subjected to.

Table 2 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)	$300 \cdot 10^6$	$N(0,0.5) \cdot 10^6$	$[-0.05,+0.05] \cdot 10^6$

In Fig. 4, measurements from the four control loops of the ALFRED are shown, on a time horizon t_M equal to 1000s: the values of the variables are kept approximately at their nominal values, at full power nominal conditions, with some measurement errors (white noise) and quantization errors.

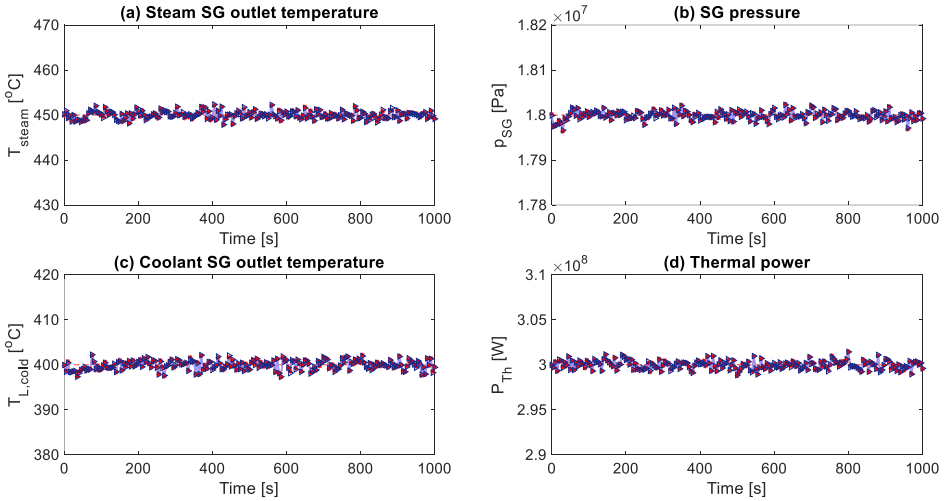

Fig. 4. Measurements from the four control loops of ALFRED at full power nominal conditions (star values for computing subsystem and triangle values for monitoring subsystem): (a) Steam SG outlet temperature; (b) SG pressure; (c) Coolant SG outlet temperature; and (d) Thermal power

2.2 Failures and cyber breaches

Both stochastic failures and cyber attacks can compromise the functionality of the ALFRED digital I&C system. Even if cyber attacks are different from components stochastic failures, they can lead to similar consequences on the system physical processes (e.g., both a stochastic failure and a cyber attack can result in sensor performance degradation [62, 63]).

To model failures and cyber attacks, a MC sampling scheme is integrated with the ALFRED model for injecting stochastic failures of sensors and cyber breaches, at

uniform random times t_R along the mission time t_M and of random magnitudes (see [15] for future details).

The occurrence of a sensor failure at random time t_R results in an altered sensor measurement $y^{sensor}(t)$, that can potentially lead the ALFRED to accidents [64-66]. Therefore, if $y(t)$ is the real value of the controlled variable y at time t, $\delta_y(t)$ is the nominal measuring error (distributed according to a normal distribution $N(0,\sigma)$), and $y^{F,sensor}(t)$ is the datastream (false measurement) when the sensor that measures $y(t)$ has failed (due to bias, drift, wider noise or freezing [15, 64, 67]):

$$y^{sensor}(t)=\begin{cases} y(t)+\delta_y(t), & t<t_R, \quad \text{normal} \\ y^{F,sensor}(t), & t\geq t_R, \quad \text{sensor failure} \end{cases} \tag{1}$$

Without loss of generality, we consider the diagnosis of the health state of the $T_{L,cold}$-PI$_3$-G_{water} control loop of Fig. 3 (all the discussion remains valid for any other control loop of the I&C system). Stochastic failures of the $T_{L,cold}$ sensor cause the measurements $T_{L,cold}^{sensor}(t)$ to differ from the real values that should be measured in the physical system due to bias, drift, wider noise and freezing [15]. Alternatively, a Denial of Service (DoS) attack can cause the blocking of a legitimate packet traffic and its substitution by a malicious packet traffic, preventing the controllers from receiving legitimate measurements and mimicking the stochastic sensor failures. Fig. 5 shows the schematics of a DoS attack, in which the computing unit is fed by a malicious packet traffic, whereas a legitimate packet traffic is fed to the monitoring unit [14, 68-72].

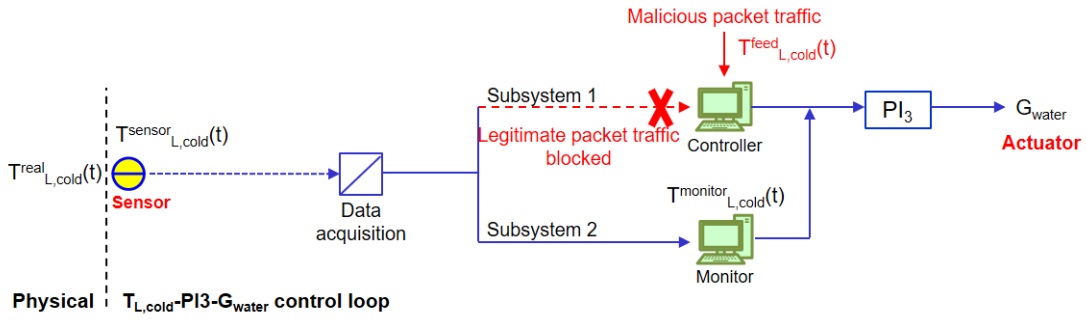

Fig. 5 Schematics of DoS attacks

2.3 The NP-CUSUM algorithm for data-driven diagnostic

Data-driven diagnostic capability based on the NP-CUSUM algorithm [14, 15] is

embedded into the control loops, for distinguishing the sensor failures from the DoS attacks. As explained in [15], the diagnostic involves two main functions: (i) reception of measurements by the controllers and feeding to the NP-CUSUM algorithm, which has been (offline) trained on different system behaviors for setting its parameters; and (ii) use of the trained algorithm and rules for discriminate recognition of failures and cyber attacks. With respect to the $T_{L,cold}$-PI$_3$-G_{water} control loop, without loss of generality:

(i) The redundant Subsystems 1 and 2 collect the measurements of $T_{L,cold}^{sensor}(t)$ at each successive time *dt*, namely, $T_{L,cold}^{feed}(t)$ and $T_{L,cold}^{monitor}(t)$, respectively, and the NP-CUSUM algorithm calculates the score function-based statistics $S_{T_{L,cold}}^{feed}(t)$ and $S_{T_{L,cold}}^{monitor}(t)$ of the collected measurements, to check whether they exceed the offline determined threshold $h_{T_{L,cold}}$: if yes, record the time(s) to alarm $\tau_{T_{L,cold}}^{feed}$ or/and $\tau_{T_{L,cold}}^{monitor}$ and proceed with the rule-based diagnostics.

(ii) If both $\tau_{T_{L,cold}}^{feed}$ and $\tau_{T_{L,cold}}^{monitor}$ are recorded (because both $S_{T_{L,cold}}^{feed}(t)$ and $S_{T_{L,cold}}^{monitor}(t)$ have exceeded the threshold), calculate the difference $\Delta\tau_{T_{L,cold}}$ between the times to alarm, $\tau_{T_{L,cold}}^{feed}$ and $\tau_{T_{L,cold}}^{monitor}$:

$$\Delta\tau_{T_{L,cold}} = \left|\tau_{T_{L,cold}}^{feed} - \tau_{T_{L,cold}}^{monitor}\right| \quad (2)$$

and compare it with a predefined reference difference $\Gamma_{T_{L,cold}}^{ref}$ for rule-based decision making:

- If $\Delta\tau_{T_{L,cold}} \leq \Gamma_{T_{L,cold}}^{ref}$, classify the event as "$T_{L,cold}$ *sensor failure*";
- If $\Delta\tau_{T_{L,cold}} > \Gamma_{T_{L,cold}}^{ref}$, classify the event as "*DoS attack*".

Notice that the NP-CUSUM algorithm requires that its parameters $\varepsilon_{T_{L,cold}}$, $h_{T_{L,cold}}$ and $\Gamma_{T_{L,cold}}^{ref}$ are customized to the different system behaviors, to guarantee the capability of discriminating between failures and cyber attacks in the $T_{L,cold}$-PI$_3$-G_{water} control loop (see [15] for future details). For illustration purpose, Fig. 6 plots $T_{L,cold}^{feed}(t)$ and

$T_{L,cold}^{monitor}(t)$ when a bias failure is injected at time t_R = 630s, with a bias factor b equal to 7.569°C, leading to $T_{L,cold}^{F,sensor}(t) = T_{L,cold}(t) + b + \delta_{T_{L,cold}}(t)$, where $t \geq t_R$. As shown in Fig. 6(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 $T_{L,cold}(t)$. Fig. 6 shows that the bias results in a very quick response of both statistics $S_{T_{L,cold}}^{feed}(t)$ and $S_{T_{L,cold}}^{monitor}(t)$, evaluated on the measurements $T_{L,cold}^{feed}(t)$ and $T_{L,cold}^{monitor}(t)$. Indeed, both statistics reach quickly the threshold $h_{T_{L,cold}}$ (dotted line) and the difference $\Delta\tau_{T_{L,cold}}$ between the times to alarm $\tau_{T_{L,cold}}^{feed}$ and $\tau_{T_{L,cold}}^{monitor}$ turns out to be actually equal to zero (i.e., less than $\Gamma_{T_{L,cold}}^{ref}$ equal to 9s) (see Fig. 6(b)), allowing for the (correct) identification of the event as a sensor failure mode and not as a cyber attack.

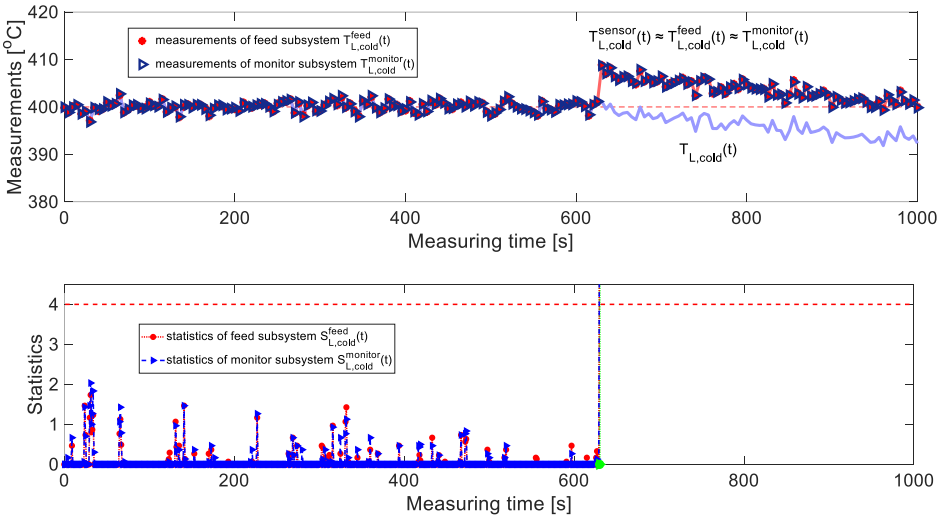

Fig. 6 $T_{L,cold}$ sensor bias failure mode: (a) the received measurements $T_{L,cold}^{feed}(t)$ and $T_{L,cold}^{monitor}(t)$ of feed and monitor Subsystems in which the bias occurs at time t_R equal to 630s; (b) the corresponding NP-CUSUM statistics $S_{L,cold}^{feed}(t)$ and $S_{L,cold}^{monitor}(t)$ for diagnosing the bias failure

Contrarily, Fig. 7(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 time to alarm difference $\Delta\tau_{T_{L,cold}}$ equal to 66s (larger than $\Gamma^{ref}_{T_{L,cold}}$) between the threshold exceedance of $S^{feed}_{T_{L,cold}}(t)$ and $S^{monitor}_{T_{L,cold}}(t)$ (see Fig. 7(b)), allowing for a (correct) identification of the event as a cyber attack.

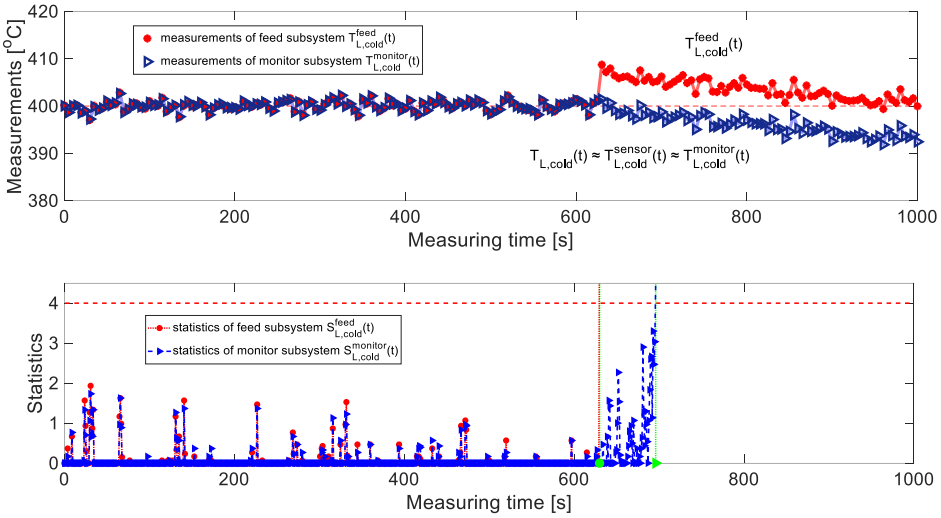

Fig. 7 Cyber attack to the computing unit mimicking a bias failure mode: (a) the received measurements $T^{feed}_{L,cold}(t)$ and $T^{monitor}_{L,cold}(t)$ of feed and monitor Subsystems in which the cyber attack occurs at time t_R equal to 630s; (b) the corresponding NP-CUSUM statistics $S^{feed}_{L,cold}(t)$ and $S^{monitor}_{L,cold}(t)$ for diagnosing the cyber attack

3. PERFORMANCE OF THE DIAGNOSTIC TOOL

The NP-CUSUM-based diagnostic is eventually interpreted by human operators, for decision-making on the action to take. The overall performance of the procedure depends on both the capability and the human operator interpretation of the diagnostic outcomes. The human cognition process for diagnostic interpretation is here modelled by BBN to structure the expert knowledge and the dependences among human factors described by Performance Shaping Factors (PSFs) [28, 30, 31, 49-53].

3.1 Human operator cognition BBN

The operator cognitive process for interpreting the diagnostic outcomes can be divided into three successive phases [29, 30, 48, 73, 74], namely: (1)

monitoring/detection (i.e., the operator observes the real-time information collected from the HMIs), (2) situation assessment (i.e., the operator develops his/her mental representation of the specific current situation) and (3) response planning (i.e., the operator takes decisions for counteracting the current situation).

When an online diagnostic outcome arrives, the operator develops his/her cognition relying on both the current understanding of the system conditions and its mental representation founded on his/her formal education, system-specific training, and operational experience, namely, the knowledge base available to the operator [75]. The operator current understanding of the real-time system observations influences his/her performance in all three phases (1), (2) and (3), whereas the mental representation responding to the specific diagnostic outcome affects his/her performance at phases (2) and (3). Besides, context variables, such as the system situation level, the human mental level and the human stress level, may impede the operator from completing the diagnostic task [30, 53, 76], as sketched in Fig. 8.

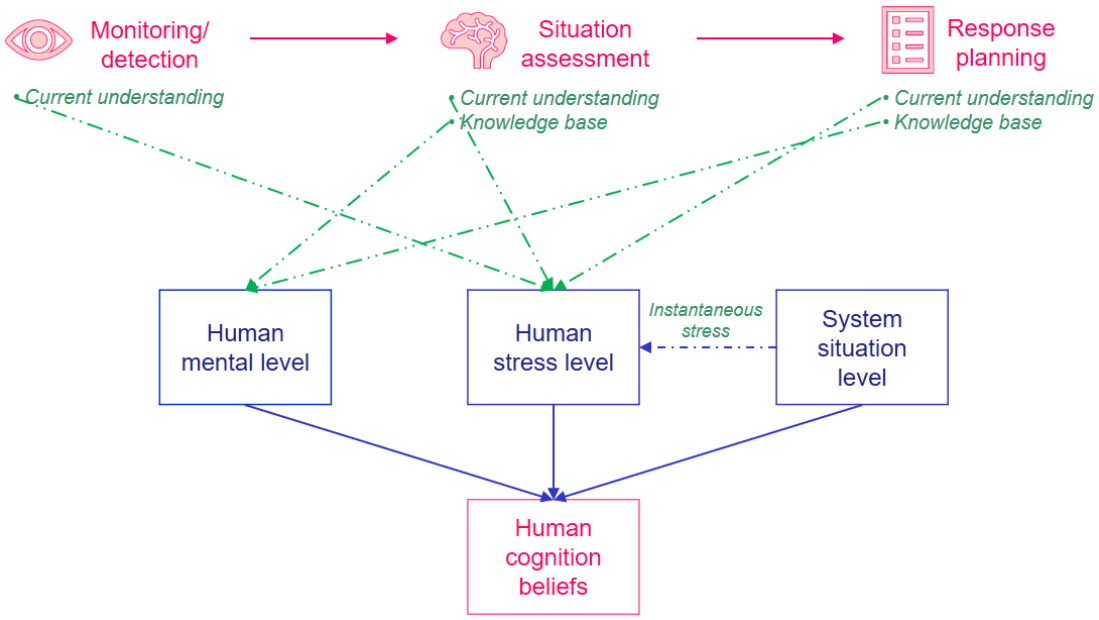

Fig. 8 The operator cognitive activity in diagnosing anomalies

To model this, the BBN of Fig. 9 contains the PSFs [30, 77] "Work process", "Diagnosis experience/training" and "Fitness of duty" (related to the operator

diagnostic knowledge base and pertaining to the human mental model), "Available diagnosis time", "Diagnosis complexity" (dependent on the states of "Diagnosis procedure" and "HMI") and "System situation level" (related to the operator understanding of the real-time system observations and, thus, identified as characteristics of the human stress model), "HMI" and "Indication of condition" (related to the system current conditions and, thus, belonging to "System situation level"). Note that the PSF "Indication of condition" is specifically introduced here for the first time, for accounting the possible failure of the human operator cognitive process in the interpretation of the data-driven diagnostic outcomes (component failures or cyber attacks).

Throughout the process, the operator performance is affected by his/her mental and stress levels depending on the diagnostic outcomes received that reflect the system situation [30, 53, 76]. Table 3 lists the PSFs with the respective levels and descriptions, whereas Fig. 9 shows the BBN model that structures, based on expert judgment [51, 78], the relationships (indicated by the arcs) among the PSFs parent nodes n_p^{α}, α=1, 2, … , 7, and the child nodes n_c^{β}, β=1, 2, … , 5, representing the operator cognitive activity, finally, determining the diagnosed state of the system as state in normal condition, under cyber attack or failed due to sensor failures.

In the BBN model, each node represents a random variable associated with discrete states, labeled as $S_p^{\alpha,\gamma}$ (for parent node) and $S_c^{\beta,\gamma}$ (for child node), hereby γ=1, 2, 3 (see Fig. 9). The parent nodes n_p^{α}, α=1, 2, … , 6, are assigned with marginal probability distributions, $p\left(S_p^{\alpha,\gamma}\middle|j\right)$ ($\sum_{\gamma=1}^{3} p\left(S_p^{\alpha,\gamma}\middle|j\right) = 1$), conditional on the operator experience to different accidental events j (i.e., sensor failure (j=a), cyber attack (j=b), or normal condition (including missed alarm) (j=c)). The NP-CUSUM data-driven diagnostic provides the operator with the current specific indication of condition (i.e., $S_p^{7,\gamma} = i$, i.e., sensor failure (i=a), cyber attack (i=b), or normal condition (including missed alarm) (i=c)), such that the marginal distribution of n_7^{α} is assigned to be $p\left(S_p^{7,\gamma} = i\right) = 1$ and $p\left(S_p^{7,\gamma} \neq i\right) = 0$, given for a specific data-driven diagnostic

outcome *i*. The relationships between nodes, namely, the probabilities of the states of the child nodes for each possible combination of its parent(s) states, are described in the form of Conditional Probability Distributions (CPDs). The CPDs for each child node are distributed in the Conditional Probability Tables (CPTs).

Once the marginal probability distributions $p\left(S_p^{\alpha,\gamma}\middle|j\right)$, α=1, 2, …, 6, γ=1, 2, 3, and the CPTs of the child nodes are assigned, the BBN model of Fig. 9 allows calculating the conditional probabilities $p\left(S_c^{1,\gamma}=k\middle|j,i\right)$ (hereafter referred to $p(k|j,i)$) of the operator diagnosing the event k to finalize as sensor failure (k=a), cyber attack (k=b), or normal condition (including missed alarm) (k=c), conditional on the combination (j,i) between the NP-CUSUM assignment i and the real accidental event j.

As discussed in [15], the NP-CUSUM algorithm may suffer from either a large false alarm rate, if the threshold is set too small (type I error), or a high missed alarm rate, if the threshold is set too large (type II error). The operator may rectify the misclassification of the data-driven diagnostic with $p(k=j|j,i\neq j)$, or erroneously respond to a correct data-driven diagnostic with $p(k\neq j|j,i=j)$ [48, 79].

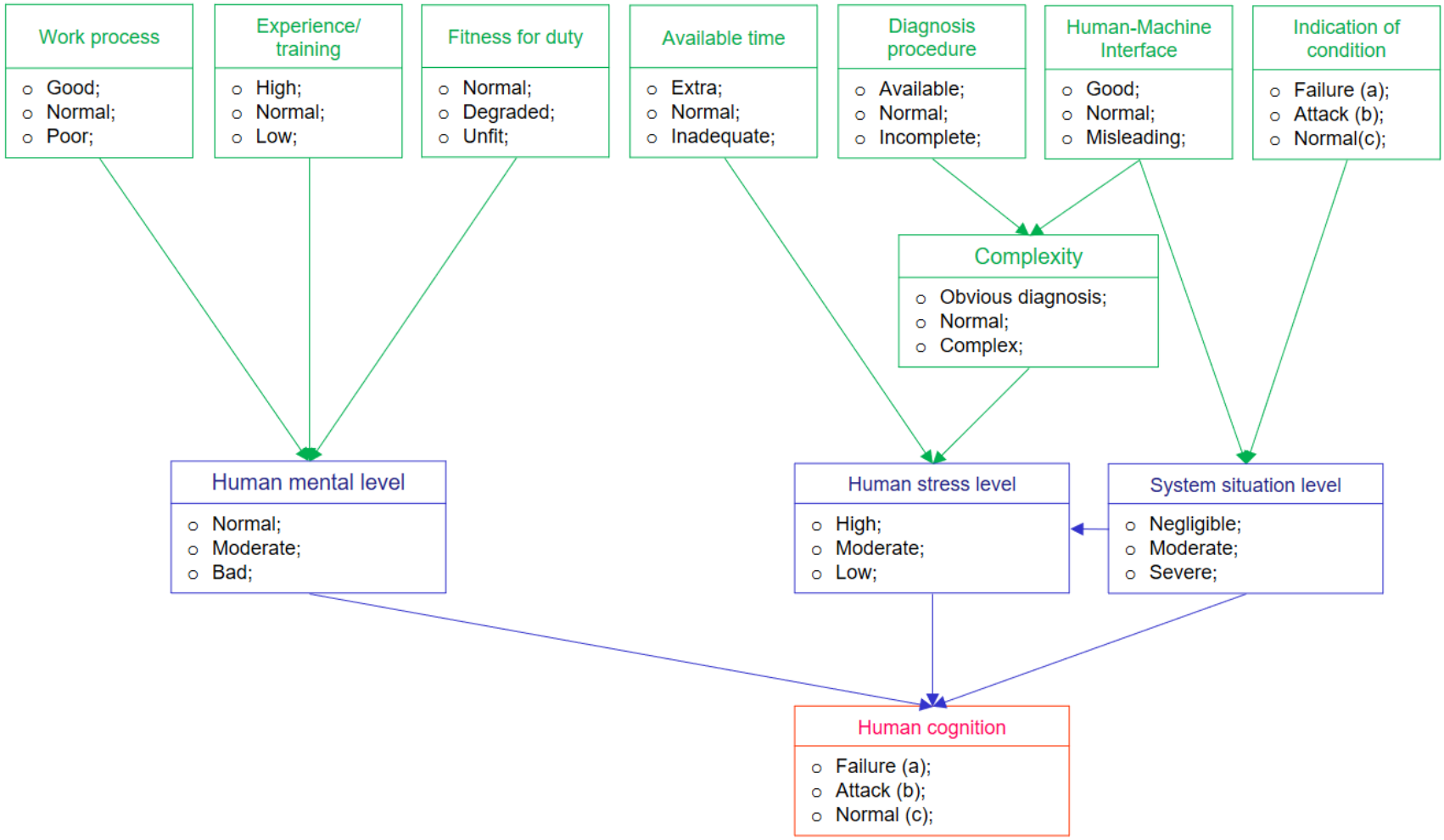

Fig. 9 The BBN model describing the human operator cognition process for final diagnostic

Table 3 PSFs affecting the human operator 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. [80-82].
			(2) Diagnosis experience/ training	High; Normal; Low.	The operator knowledge base, experience and training related to the diagnostic task [82].
			(3) Fitness of duty	Normal; Degraded; Unfit.	The operator physical and mental fitness to perform the diagnosis task at the time [80, 82].
	(3) System situation level	/	(4) Available diagnosis time	Extra; Normal; Inadequate.	The operator available time to diagnose an abnormal event [82].
		(5) Diagnosis complexity	(5) Diagnosis procedure	Available; Normal; Incomplete.	The existence of feasible procedures for the diagnosis and response planning tasks [82, 83].
			(6) Human-Machine Interface (HMI)	Good; Normal; Misleading.	The availability of real-time physical information from Human-Machine Interfaces (HMIs) for the operator to carry out the diagnostic task [82].
	(4) Human stress level		(6) HMI	Good; Normal; Misleading.	The availability of real-time physical information from HMIs for the operator to carry out the diagnostic task [82].

			(7) Indication of condition	*i*=a;
				i=b;
				i=c.
					The clarity of the data-driven diagnostic indications that assist the operator in diagnosing the anomaly [84].

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;

3.2 Overall diagnostic performance

With reference to a recorded event *j*, the human operator correctly diagnoses the event when his/her assignment *k* is consistent with *j*, even if the data-driven diagnostic indicates a misclassified system state *i*. Vice versa, if the operator indication *k* is not consistent with the event *j*, an incorrect diagnostic of the accidental event is made by the operator, regardless of the correctness of the data-driven assignment *i*.

Table 4 summarizes all possible human operator assignments *k* of the event *j*, with respect to the different data-driven diagnostic assignments *i*. In the Table, conditional on the assignment *i*, the correct diagnostic of the event is tagged by the symbol "√" (see Column 4), with a conditional probability equal to $p(j, k = j|i)$, where *i*, *j*, *k* = a, b or c (see Column 5), whereas, the incorrect diagnostic of the event is tagged by the symbol "×", with a conditional probability equal to $p(j, k \neq j|i)$, where *i*, *j*, *k* = a, b or c (see Column 5).

In the end, only the consistency of the human operator assignment *k* with the event *j* gives a correct diagnostic: then, the probability of correct diagnostic p^i_{correct}, conditional on the data-driven diagnostic *i*, is obtained by summing the probabilities of the occurrences tagged by "√" in Table 4.

$$p^i_{\text{correct}} = \sum_{j=k=a}^{c} p(j, k = j|i) \tag{3}$$

According to the chain rule of conditional probability [85], Eq. (3) can change to:

$$p^i_{\text{correct}} = \sum_{j=k=a}^{c} p(k = j|j, i) \cdot p(j|i) \tag{4}$$

where $p(j|i)$ is the probability that the event is *j*, when the data-driven diagnostic is *i* (i.e., the probability of correct diagnosis of the data-driven algorithm if $j = i$). This represents the performance of the data-driven diagnostic, which, as discussed in [15], can be empirically estimated from N_v tests of (unknown) failures and cyber attacks. On the other hand, $p(k = j|j, i)$ is the ability of the operator to interpret the diagnostic outcome *i* and correctly assign his/her diagnostic *k* consistent with the occurred event *j*, i.e., *k*=*j*.

Table 4 All possible diagnostic assignments

Data-driven diagnostic, *i*	Human operator assignment, *k*	Event, *j*	Tag	Conditional probability
Sensor failure (a)	a	a	√	p(j=a, k=a | i=a)
		b	×	p(j=b, k=a | i=a)
		c	×	p(j=c, k=a | i=a)
	b	a	×	p(j=a, k=b | i=a)
		b	√	p(j=b, k=b | i=a)
		c	×	p(j=c, k=b | i=a)
	c	a	×	p(j=a, k=c | i=a)
		b	×	p(j=b, k=c | i=a)
		c	√	p(j=c, k=c | i=a)
Cyber attack (b)	a	a	√	p(j=a, k=a | i=b)
		b	×	p(j=b, k=a | i=b)
		c	×	p(j=c, k=a | i=b)
	b	a	×	p(j=a, k=b | i=b)
		b	√	p(j=b, k=b | i=b)
		c	×	p(j=c, k=b | i=b)
	c	a	×	p(j=a, k=c | i=b)
		b	×	p(j=b, k=c | i=b)
		c	√	p(j=c, k=c | i=b)
Normal condition (no indication) (c)	a	a	√	p(j=a, k=a | i=c)
		b	×	p(j=b, k=a | i=c)
		c	×	p(j=c, k=a | i=c)
	b	a	×	p(j=a, k=b | i=c)
		b	√	p(j=b, k=b | i=c)
		c	×	p(j=c, k=b | i=c)
	c	a	×	p(j=a, k=c | i=c)
		b	×	p(j=b, k=c | i=c)
		c	√	p(j=c, k=c | i=c)

Notes:
1) hereafter "a" refers to "sensor failure", "b" refers to "cyber attack", and "c" refers to "normal condition" (including missed alarm);
2) " √ " refers to correct diagnostic, and "×" refers to incorrect diagnostic.

To practically calculate the overall performance p^{i}_{correct} for different data-driven diagnostic *i*, a general MC approach (sketched in Fig. 10) is proposed in line with [52, 86] for populating the CPTs at the child nodes of the operator BBN. We proceed as follows:

At the *m*-th MC run, $m = 1, 2, \ldots, N_m$:

(1) Set the distributions of PSFs for each event $j = a$, b or c (see Appendix A), i.e., the operator knowledge and experience, relative to the class of events *j*;

(2) Set $i = a$, b or c, and the evidence of the parent node $s_p^{7,\gamma}$ as equal to *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 $s_p^{7,\gamma}$) 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 use of the five-step functional interpolation method [52, 86]:

(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 B);

(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) Assign the normal distribution $N\left(u\left(e_c^{\beta,\theta}\right),\sigma\left(e_c^{\beta,\theta}\right)\right)$ to the θ-th element of the β-th child node n_c^{β} CPT, β=1, 2, … , 5, and assign the states $s_c^{\beta,\gamma}$ of the child node n_c^{β} with the $N\left(u\left(e_c^{\beta,\theta}\right),\sigma\left(e_c^{\beta,\theta}\right)\right)$ pdf values at the $s_c^{\beta,\gamma}$ states anchor values (i.e., γ assigned equal to 1, 2 (and 3), respectively (see Appendix B)), being the conditional probability scales of $s_c^{\beta,\gamma}$ in $e_c^{\beta,\theta}$, i.e., $\eta\left(s_c^{\beta,\gamma}\middle|e_c^{\beta,\theta}\right)$;

(4d) Normalize $\sum_{\gamma}\eta\left(s_c^{\beta,\gamma}\middle|e_c^{\beta,\theta}\right)$ to 1, leading the scale values to being the conditional probabilities of the $s_c^{\beta,\gamma}$ states, i.e., $p\left(s_c^{\beta,\gamma}\middle|e_c^{\beta,\theta}\right)$, in the θ-th element of the child node n_c^{β} CPT;

(4e) Collect the CPDs of all the elements and, build the CPTs for the n_c^{β} 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

$p_m(k = j|j,i)$ conditional on the combination (*j*,*i*) with current assigned values of *j* and *i*;

(6) Repeat steps (1) to (5), and collect the estimates of $p_m\left(k = j \mid j,i\right)$ for the nine combinations (*j*,*i*): (*a*,*a*), (*a*,*b*), (*a*,*c*), (*b*,*a*), (*b*,*b*), (*b*,*c*), (*c*,*a*), (*c*,*b*) and (*c*,*c*);

(7) Feed $p_m(k = j|j,i)$ and the tested $p(j|i)$ values to Eq. (4), to obtain the estimates of the performance $p^i_{\text{correct},m}$, with respect to the different data-driven diagnostic indications *i*.

Repeat steps (1) to (7) for N_m times, and obtain the confidence intervals of the p^i_{correct}, with respect to different data-driven diagnostic *i*.

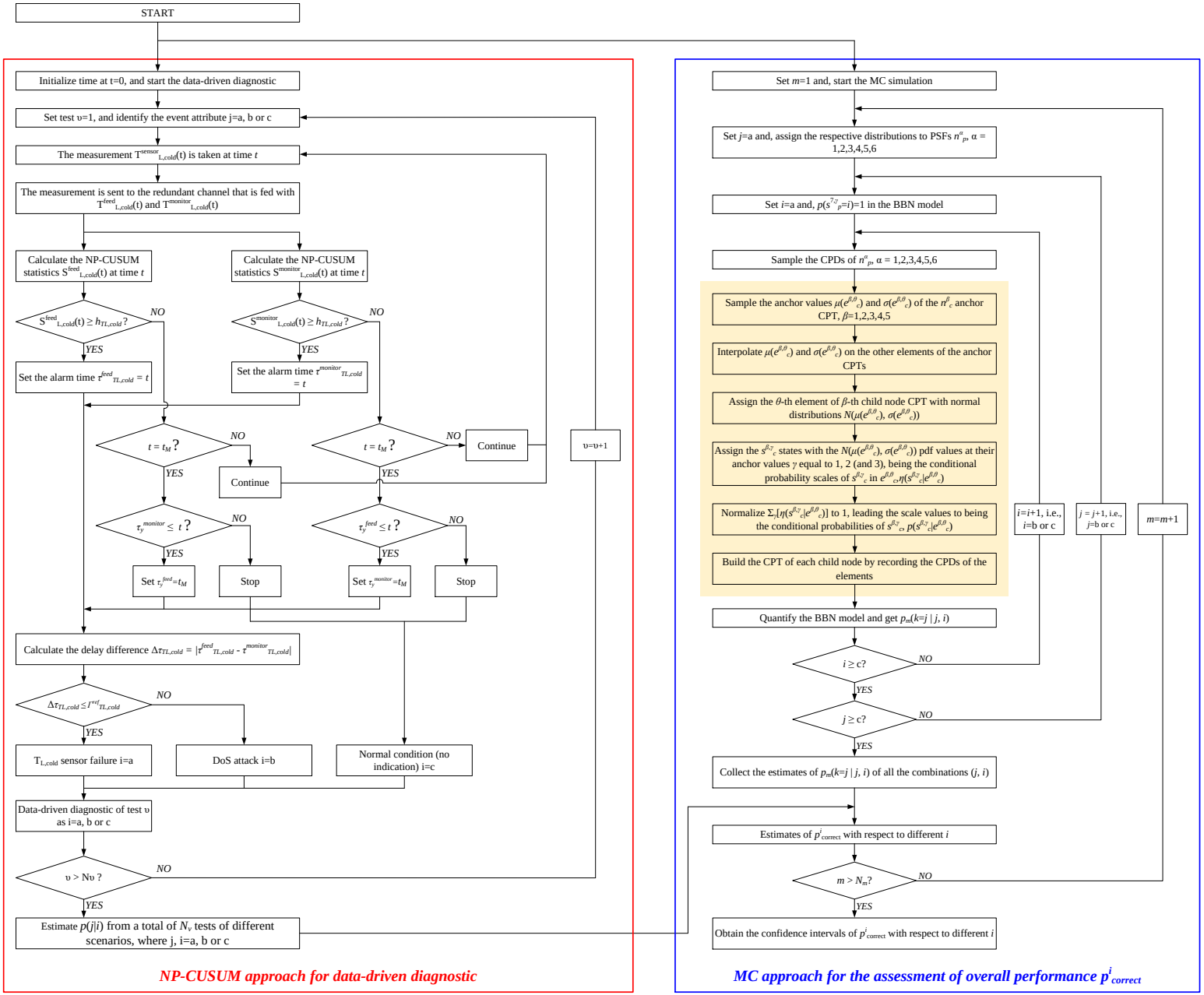

Fig. 10 The flowchart for estimating the diagnostic performance

4. RESULTS

We generated 389 $T_{L,cold}$ transients due to sensor failures, 392 transients due to DoS attacks and 219 transients of normal operation scenarios out of a total of $N_v = 1000$ tests scenarios of ALFRED evolution. 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 the NP-CUSUM parameters randomly sampled from their distributions listed in Table 5.

Table 5 Parameters of the NP-CUSUM algorithm

Parameter	Description	Distribution	Unit
$\varepsilon_{T_{L,cold}}$	The NP-CUSUM tuning parameter	U(2,5)·10^5	-
$h_{T_{L,cold}}$	The NP-CUSUM positive threshold	U(3.8,4.0)	-
$\Gamma^{ref}_{T_{L,cold}}$	The reference delay difference between $\tau^{feed}_{T_{L,cold}}$ and $\tau^{monitor}_{T_{L,cold}}$	U(8,9)	s

Table 6 collects the number of the data-driven diagnostic outputs, and lists the estimates of $p(j|i)$: the data-driven diagnostic classifies the N_v tests into 386 sensor failures (a), 386 DoS attacks (b) and 228 normal condition (c), resulting in probabilities of correct assignment $p(j = i|i)$ equal to 0.9611, 0.9819 and 0.8772, respectively. It is worth noting that $p(j = c|i = c)$ is smaller than $p(j = a|i = a)$ and $p(j = b|i = b)$, because the NP-CUSUM algorithm suffers of a relatively high missed alarm rate when the occurring events negligibly affect the controlled variables and the system functionality.

Table 6 Performance of the NP-CUSUM diagnostic

NP-CUSUM diagnostic *i*	Number of events	Correctness	*p*(*j* | *i*)	Probability
$T_{L,cold}$ sensor failure (a)	386	*Correct*	*p*(*j*=a | *i*=a)	371/386 (0.961)
		Misclassfication of cyber attack	*p*(*j*=b | *i*=a)	1/386
		Misclassfication of normal condition	*p*(*j*=c | *i*=a)	14/386
DoS attack (b)	386	Misclassfication of component failure	*p*(*j*=a | *i*=b)	2/386
		Correct	*p*(*j*=b | *i*=b)	379/386 (0.982)
		Misclassfication of normal condition	*p*(*j*=c | *i*=b)	5/386
normal condition (c)	228	Misclassfication of component failure	*p*(*j*=a | *i*=c)	16/228
		Misclassfication of cyber attack	*p*(*j*=b | *i*=c)	12/228
		Correct	*p*(*j*=c | *i*=c)	200/228 (0.877)

The operator cognitive errors in interpreting the diagnostic outcomes have been

calculated as discussed in Section 3.2, by running the operator cognition BBN of Fig. 9: the correct diagnostic probability p^i_{correct}, given data-driven diagnostic i (=a, b or c) is calculated according to Eq. (4), after N_m=1000 runs, along with the double-sided 95% confidence intervals of p^i_{correct}. As shown in Fig. 11, the mean values of p^i_{correct} (circles in Fig. 11) turn out to be equal to 0.966, 0.923 and 0.943, with respect to the different data-driven diagnostic i (i.e., $T_{L,cold}$ sensor failures (a), DoS attacks (b) and normal conditions including missed alarms (c), respectively).

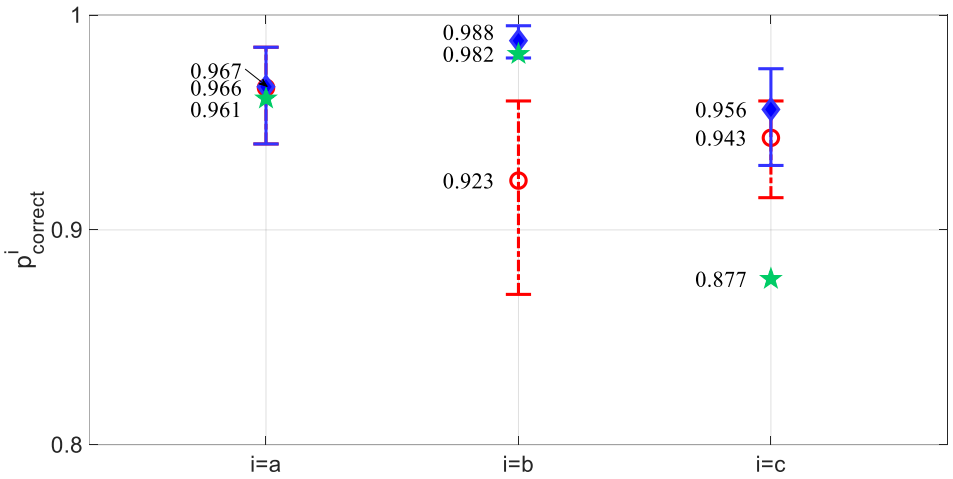

Fig. 11 Estimates of the double-sided 95% confidence intervals of the correct diagnostic probabilities

The results of Fig. 11 show that the mean values of p^a_{correct} (equal to 0.966) and p^c_{correct} (equal to 0.943) are respectively larger than $p(j = a|i = a)$ (equal to 0.961) and $p(j = c|i = c)$ (equal to 0.877) (stars in Fig. 11). This shows that in these cases the operator expertise increases the performance of the data-driven algorithm by correcting events that were misclassified by the NP-CUSUM. The confidence interval of p^b_{correct} turns out to be large and its mean value (equal to 0.923) turns out to be smaller than the performance of the data-driven diagnostic tool ($p(j = b|i = b)$ equal to 0.982 labeled as star in Fig. 11). Conversely, this shows that, in this case, the lack of operators experience in correctly interpreting the NP-CUSUM outcome results in mistaking the diagnostic and in worsening the diagnosis performance with respect to cyber attacks.

Furthermore, we have repeated the analysis by running N_m=1000 runs of the BBN,

assuming a fully skilled operator with respect to diagnosing cyber attack events, i.e., $p\left(s_p^{2,1}\middle|i=b\right)=1$. Double-sided 95% confidence intervals of the correct diagnostic probabilities p_{correct}^{i} with respect to the different data-driven diagnostics *i* (i.e., *a*, *b* and *c*, respectively) are shown in solid lines in Fig. 11: the confidence interval of p_{correct}^{b} turns out to be narrower and its mean value (equal to 0.988) (diamond in Fig. 11) turns out to be larger than the performance of the data-driven diagnostic (equal to 0.982), as expected.

5. CONCLUSIONS

In this study, we have proposed a framework for the discrimination analysis of cyber attacks and component failures in Cyber-Physical Systems (CPSs). The framework combines, for the first time, a data-driven diagnostic approach (Non-Parametric CUmulative SUM (NP-CUSUM), in this case) with a Bayesian Belief Network (BBN) that is originally tailored to model the human operator cognition process of interpreting the diagnostic outcomes. The BBN is used to structure the expert knowledge and other factors (in particular, the indication of the data-driven diagnostic outcomes) that influence the human cognition process for diagnostic.

In the application, the work contributes to the process of enabling the interaction between data-driven diagnostic systems and human operators actions for supporting operator decisions with respect to cyber attacks in CPSs, with the aim of reducing false alarms, missed alarms, or misclassifications of cyber attacks as components failures, and vice versa.

We have illustrated the work considering the digital Instrumentation and Control (I&C) system of the Advanced Lead-cooled Fast Reactor European Demonstrator (ALFRED). The results of the case study show that the proposed diagnostic approach is capable of identifying most of the generated failure/attack scenarios, with low frequency of misclassifications, and that, in the case considered, the operator increases the diagnosis performance for sensors failures, but not for cyber attacks. The results persuasively demonstrate that cyber attacks are less diagnosable compared to

components failures.

A simplistic assumption is done that cyber attacks occur at random times from the viewpoint of the defender; in reality, attacker is likely to launch cyber attacks at preferred times when his/her objectives can be maximized and the investment minimized. This challenges the diagnostic task, and therefore, with due caution, future work will regard the role of the attacker decision making (e.g., based on a cost benefit analysis) for choosing the optimal time of attack.

APPENDIX A

With respect to the events of sensor failure (j = a) and normal conditions (including missed alarms, j = c), the probability distributions of the PSF states $p\left(s_p^{\alpha,\gamma}\middle|j\right)$, $\alpha = 1, 2, \cdots, 6$, $\gamma = 1, 2, 3$, are taken as uniform (see Table A1), according to expert judgment, whose mean values are given [87]. Under DoS attack events (j = b), the operator is assumed to be less experienced (n_p^2) and the diagnosis procedure (n_p^5) relatively incomplete, such that the probability distributions $p\left(s_p^{2,\gamma}\middle|j = b\right)$ and $p\left(s_p^{5,\gamma}\middle|j = b\right)$ result in those in the last column of Table A1.

It us worth mentioning that, with respect to the MC simulation presented in Section 3.2, the sampled values from the distributions of $p\left(s_p^{\alpha,\gamma}\middle|j\right)$, $\gamma = 1, 2, 3$, for each parent node n_p^α, given an event j, are normalized to the sum equal to 1 (i.e., $\sum_{\gamma=1}^{3} p\left(s_p^{\alpha,\gamma}\middle|j\right) = 1$, given an α and a j).

Table A1 Identification of probability distributions for the states of PSFs

Parent nodes, n_p^{α}, α=	States, $S_p^{\alpha,\gamma}$, γ=	$p(s_p^{\alpha,\gamma}\|j)$, when j = a or c	$p(s_p^{\alpha,\gamma}\|j)$, when 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 B

As suggested in [52], we build the anchor CPTs for the child nodes n_c^{β} (β = 1, 2, 3, 4, 5) of the BBN model of Fig. 9, as listed in Tables A2 to A6, 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 anchor values equal to 1, 2 (and 3) for identifying the corresponding CPD scales (i.e., pdf values at the anchor values equal to 1, 2 (and 3)), once the uniform distributions at all the elements of the anchor CPTs are generated.

Table A2 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 $s_c^{1,\gamma}$, correct (i.e., k=j) and incorrect (i.e., $k \neq j$) diagnostic are assigned with the anchor values γ equal to 1 and 2, respectively.

Table A3 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 $s_c^{2,\gamma}$, Normal, Moderate and Bad are assigned with the anchor values γ equal to 1, 2 and 3, respectively.

Table A4 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 $s_c^{3,\gamma}$, Low, Moderate and High are assigned with the anchor values γ equal to 1, 2 and 3, respectively.

Table A5 The anchor CPT of n_c^4 (System situation level)

Indication of condition		$i = j$ (e.g., a)	$i \neq j$ (b)	$i \neq j$ (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 $s_c^{4,\gamma}$, Negligible, Moderate and Severe are assigned with the anchor values γ equal to 1, 2 and 3, respectively.

Table A6 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 $s_c^{5,\gamma}$, Obvious, Normal and Complex are assigned with the anchor values γ equal to 1, 2 and 3, respectively.

REFERENCES

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

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

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

[4] 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.

[5] 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.

[6] Zio E. The future of risk assessment. Reliability Engineering & System Safety. 2018;177:176-90.

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

[8] 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.

[9] Hausken K, Levitin G. Review of systems defense and attack models. International Journal of Performability Engineering. 2012;8:355-66.

[10] Wang W, Di Maio F, Zio E. Adversarial Risk Analysis to Allocate Optimal Defense Resources for Protecting Cyber–Physical Systems from Cyber Attacks. Risk Analysis. 2019.

[11] Levitin G, Hausken K. Parallel systems under two sequential attacks. Reliability Engineering & System Safety. 2009;94:763-72.

[12] Piccinelli R, Sansavini G, Lucchetti R, Zio E. A general framework for the assessment of power system vulnerability to malicious attacks. Risk Analysis. 2017;37:2182-90.

[13] Mo H, Sansavini G. Dynamic defense resource allocation for minimizing unsupplied demand in cyber-physical systems against uncertain attacks. IEEE Transactions on Reliability. 2017;66:1253-65.

[14] 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.

[15] 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.

[16] 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.

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

[18] 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.

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

[20] Roberts SW. Control Chart Tests Based on Geometric Moving Averages. Technometrics. 1959;42:97-101.

[21] Ozay M, Esnaola I, Vural FTY, Kulkarni SR, Poor HV. Machine learning methods for attack detection in the smart grid. IEEE transactions on neural networks and learning systems. 2016;27:1773-86.

[22] Buczak AL, Guven E. A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Communications Surveys & Tutorials. 2016;18:1153-76.

[23] Tan Z, Jamdagni A, He X, Nanda P, Liu RP. A system for denial-of-service attack detection based on multivariate correlation analysis. IEEE transactions on parallel and distributed systems. 2014;25:447-56.

[24] Li Y, Quevedo DE, Dey S, Shi L. SINR-based DoS attack on remote state estimation: A game-theoretic approach. IEEE Transactions on Control of Network Systems. 2017;4:632-42.

[25] Xiao L, Li Y, Han G, Liu G, Zhuang W. PHY-layer spoofing detection with reinforcement learning in wireless networks. IEEE Transactions on Vehicular Technology. 2016;65:10037-47.

[26] Xiao L, Wan X, Lu X, Zhang Y, Wu D. IoT Security Techniques Based on Machine Learning: How Do IoT Devices Use AI to Enhance Security? IEEE Signal Processing Magazine. 2018;35:41-9.

[27] Zhang J, Blum RS, Poor HV. Approaches to Secure Inference in the Internet of Things: Performance Bounds, Algorithms, and Effective Attacks on IoT Sensor Networks. IEEE Signal Processing Magazine. 2018;35:50-63.

[28] Jiang J, Wang Y, Zhang L, Wu D, Li M, Xie T, et al. A cognitive reliability model research for complex digital human-computer interface of industrial system. Safety science. 2018;108:196-202.

[29] 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.

[30] 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.

[31] 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.

[32] Gratian M, Bandi S, Cukier M, Dykstra J, Ginther A. Correlating Human Traits and Cybersecurity Behavior Intentions. Computers & Security. 2017;73:345–58.

[33] 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.

[34] Commission USNR. Technical Basis and Implementation Guidelines for A Technique for Human Event Analysis (ATHEANA) Washington, DC 20555-0001 2000.

[35] Embrey D, Humphreys P, Rosa E, Kirwan B, Rea K. SLIM-MAUD: an approach to assessing human error probabilities using structured expert judgment. Volume II. Detailed analysis of the technical issues. Brookhaven National Lab.; 1984.

[36] Seaver DA, Stillwell WG. Procedures for using expert judgment to estimate human-error probabilities in nuclear power plant operations.[PWR; BWR]. Decision Science Consortium, Inc., Falls Church, VA (USA); 1983.

[37] Cacciabue PC, Decortis F, Drozdowicz B, Masson M, Nordvik J-P. COSIMO: a cognitive simulation model of human decision making and behavior in accident management of complex plants. IEEE Transactions on Systems, Man, and Cybernetics. 1992;22:1058-74.

[38] Fan X, Chen P-C, Yen J. Learning HMM-based cognitive load models for supporting human-agent teamwork. Cognitive Systems Research. 2010;11:108-19.

[39] Buczak AL, Guven E. A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Communications Surveys & Tutorials. 2015;18:1153-76.

[40] Tan Z, Jamdagni A, He X, Nanda P, Liu RP. A system for denial-of-service attack detection based on multivariate correlation analysis. IEEE transactions on parallel and distributed systems. 2013;25:447-56.

[41] Li SY, Tang LC, Ng SH. Nonparametric CUSUM and EWMA Control Charts for Detecting Mean Shifts. Journal of Quality Technology. 2010;42:209-26.

[42] Yang SF, Cheng SW. A new non-parametric CUSUM mean chart. Quality & Reliability Engineering International. 2011;27:867-75.

[43] Alippi C, Boracchi G, Roveri M. Hierarchical change-detection tests. IEEE transactions on neural networks and learning systems. 2017;28:246-58.

[44] 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.

[45] Sorrells C, Qian L, Li H. Quickest detection of denial-of-service attacks in cognitive wireless networks. Homeland Security2012. p. 580-4.

[46] 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.

[47] 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.

[48] Commission UNR. Technical basis and implementation guidelines for a technique for human event analysis (ATHEANA). NUREG-1624, Rev. 2000;1.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] Skogestad S, Postlethwaite I. Multivariable feedback control: analysis and design: Wiley New York; 2007.

[58] Grasso G, Petrovich C, Mikityuk K, Mattioli D, Manni F, Gugiu D. Demonstrating the effectiveness of the European LFR concept: the ALFRED core design. International Conference on FAST Reactors and Related Fuel Cycles: Safe Technologies and Sustainable Scenarios2015.

[59] 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.

[60] Gray RM, Neuhoff DL. Quantization. IEEE Transactions on Information Theory. 1998;44:2325-83.

[61] Widrow B. Analysis of amplitude-quantized sampled-data systems. Electrical Engineering. 1961;80:450-.

[62] 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.

[63] 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.

[64] Boskvic JD, Mehra RK. Stable adaptive multiple model-based control design for accommodation of sensor failures. 2002;3:2046-51 vol.3.

[65] Jones HL. Failure detection in linear systems: Massachusetts Institute of Technology; 1973.

[66] 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.

[67] Wang W, Maio FD, 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.

[68] 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.

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

[70] Yuan Y, Yuan H, Guo L, Yang H, Sun S. Resilient Control of Networked Control System Under DoS Attacks: A Unified Game Approach. IEEE Transactions on Industrial Informatics. 2016;12:1786-94.

[71] 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.

[72] Wang W, Cammi A, Maio FD, 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.

[73] Naderpour M, Lu J, Zhang G. A human-system interface risk assessment method based on mental models. Safety science. 2015;79:286-97.

[74] 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.

[75] O'Hara JM. The effects of interface management tasks on crew performance and safety in complex, computer-based systems: US Nuclear Regulatory Commission, Office of Nuclear Regulatory Research …; 2002.

[76] 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.

[77] Zwirglmaier K, Straub D, Groth KM. Capturing cognitive causal paths in human reliability analysis with Bayesian network models. Reliability Engineering & System Safety. 2017;158:117-29.

[78] 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.

[79] 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.

[80] 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.

[81] 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.

[82] Gertman D, Blackman H, Marble J, Byers J, Smith C. The SPAR-H human reliability analysis method. US Nuclear Regulatory Commission. 2005.

[83] 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.

[84] 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.

[85] Samudrala R, Moult J. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction1. Journal of molecular biology. 1998;275:895-916.

[86] Podofillini L, Mkrtchyan L, Dang V. Aggregating expert-elicited error probabilities to build HRA models. Safety and Reliability: Methodology and Applications: CRC Press; 2014. p. 1119-28.

[87] 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.