INL/JOU-22-70031-Revision-0

Trustworthiness modeling and evaluation for a nearly autonomous management and control system

May 2024

Longcong Wang, Nam Dinh, Linyu Lin

Changing the World's Energy Future

Idaho National Laboratory

INL/JOU-22-70031-Revision-0

Trustworthiness modeling and evaluation for a nearly autonomous management and control system

Longcong Wang, Nam Dinh, Linyu Lin

May 2024

Idaho National Laboratory
Idaho Falls, Idaho 83415

http://www.inl.gov

Prepared for the
U.S. Department of Energy
Under DOE Idaho Operations Office
Contract DE-AC07-05ID14517

Trustworthiness Modeling and Evaluation for a Nearly Autonomous Management and Control System

Longcong Wang[a,*], Linyu Lin[b], Nam Dinh[a]

[a]*Department of Nuclear Engineering, North Carolina State University, 2500 Stinson Dr, 27695, Raleigh, NC, United States*

[b]*Instrumentation, Controls, and Data Science, Idaho National Laboratory, 83415, Idaho Falls, ID, United States*

Abstract

The Nearly Autonomous Management and Control (NAMAC) system supports the advanced reactor operation by recommending control actions to operators based on real-time measurements and digital twins (DTs) learning from the knowledge base. To enable the safe and reliable use of autonomous technologies, NAMAC and its recommendations should be trustworthy to operators and regulators at both the design and operation stages. This study proposes a NAMAC trustworthiness modeling and evaluation framework supported by trustworthiness ontologies and evidence-based approaches. The development-time and run-time ontologies are separately constructed and then converted to Bayesian networks to quantitatively evaluate the NAMAC trustworthiness. This evaluation is demonstrated by collecting and characterizing evidence from NAMAC practices, such as the development and assessment of the NAMAC system, data coverage assessment, and the training and optimizations of neural-network-based DTs. Our proposed approach can aggregate various trustworthiness attributes of complex artificial-intelligence-supported systems for safety-critical applications. It also considers the interaction between different DTs and extends beyond the trustworthiness evaluation of a single DT. The evidence-based method enhances the transparency of the trustworthiness modeling and evaluation processes and helps identify uncertainties and subjectivity involved in the processes.

Keywords: Trustworthiness, Ontology, Digital Twin, Data Coverage Assessment

*Corresponding author

Email addresses: longcwang3@gmail.com (Longcong Wang), linyu.lin@inl.gov (Linyu Lin), ntdinh@ncsu.edu (Nam Dinh)

1. Introduction

The Nearly Autonomous Management and Control (NAMAC) system is developed to provide recommendations to advanced reactor operators during nominal and accidental conditions [1, 2]. It derives such recommendations based on current reactor states and forecast of reactor future. Different digital twins (DTs) supported by machine learning (ML) algorithms extract and store the required knowledge, and those DTs are integrated through a structured workflow. As the NAMAC system intends to support the autonomous control of safety-critical systems, one of the essential requirements is that NAMAC and its recommendations be trustworthy. Based on the existing definitions of trustworthiness reviewed in [3], this study defines NAMAC trustworthiness as *an assurance that the NAMAC system will work as expected*.

For complex engineering systems, such as cyber-physical systems (CPSs) and autonomous systems, trustworthiness has been identified as one of the crucial facets that need to be analyzed [3–10]. Mohammadi [3] proposed a CPS trustworthiness management framework, which includes key trustworthiness concepts, trustworthiness evaluations, and run-time trustworthiness maintenance. In the CPS framework (CPSF) proposed by the U.S. National Institute of Standards and Technology (NIST) [5], trustworthiness was identified as one of the nine aspects of a CPS and considered a collection of conceptually related concerns. Based on NIST's CPSF, CPS trustworthiness ontologies were developed to represent and reason the CPS trustworthiness [6, 7, 11]. For smart grids, Gao et al. [8] defined and evaluated trustworthiness based on three metrics: availability, reliability, and integrity. Ding et al. [9] handled software trustworthiness evaluation as a multiple-attribute decision analysis and decomposed the software trustworthiness evaluation indicator set into several trustworthiness attributes, such as reliability and safety. Cho et al. [10] treated trustworthiness as a system-level metric consisting of four key aspects: security, trust, resilience, and agility. Overall, we may conclude that trustworthiness is usually considered a multiattribute concept.

As the volume of research on artificial intelligence (AI) and ML in engineering applications rapidly grows, trustworthiness of these AI/ML models and systems is gaining significant attention. The United States Nuclear Regulatory Commission (U.S. NRC) identified AI trustworthiness as a

potential technical consideration and key research area for regulatory decision-making [12]. NIST developed an AI risk management framework to prompt AI trustworthiness and articulate characteristics of a trustworthy AI [13]. Chelouati et al. [14] explored the capability of goal structuring notation for building a safety case for autonomous trains, which hierarchically distinguished AI-based components from AI techniques. However, only high-level argument pattern is presented in [14]. Assessing and enhancing the trustworthiness of a single AI/ML model via uncertainty quantification (UQ) has been widely used in machine fault diagnosis [15–17], prognosis health management [18], and risk predictions [19]. Uncertainty, especially epistemic uncertainty, has been used as a trustworthiness metric for domain shifting and data coverage issues [15–17, 20]. However, other trustworthiness characteristics, including data adequacy and DT training and assessment results are not explicitly included and could deteriorate the final trustworthiness. Moreover, those studies focus on single AI/ML models, while the modeling and evaluation of the trustworthiness of a system with a network of AI/ML models have not been investigated. Lastly, the trustworthiness assessment approach in this work considers the evidence gathered from both development time and run time, which accounts for the dynamic features of AI/ML trustworthiness and continuously monitors its performance at run time.

Although this is the first time NAMAC trustworthiness is formally modeled and evaluated, efforts were made during the NAMAC development to monitor and enhance its trustworthiness [2, 20–23]. The discrepancy checker [2, 21] has been designed as an essential part of the NAMAC operational workflow to prevent major failures during NAMAC operation. Meanwhile, the NAMAC functionality relies on several DTs, and numerous techniques are implemented to enhance DT quality [2, 22, 23], such as physics-guided ML and Bayesian-based hyperparameter tuning. During NAMAC operation, the data coverage assessment (DCA) [20] is designed to detect unseen data, which can cause the failure of data-driven DTs. When NAMAC trustworthiness is systematically evaluated, these existing efforts can provide the required evidence. However, the evidence can be highly heterogeneous, and how the heterogeneous evidence can be characterized and integrated to evaluate NAMAC trustworthiness needs further investigation.

This study proposes and demonstrates an evidence-based framework for modeling and evaluating NAMAC trustworthiness. We first model NAMAC trustworthiness by identifying related

attributes. Next, trustworthiness is evaluated based on evidence of each attribute. Section 2 provides the technical background for the NAMAC system and work related to its trustworthiness modeling, which narrows the research scope. Section 3 presents the overall methodology for NAMAC trustworthiness modeling and evaluation. Section 4 demonstrates how development-time NAMAC trustworthiness can be evaluated by characterizing and integrating evidence from existing works. Section 5 shows how run-time trustworthiness can be evaluated with given DCA and discrepancy checker results.

2. Technical Background

2.1. NAMAC system

2.1.1. Simplified NAMAC operational workflow

As shown in Figure A.16, the complete NAMAC operational workflow consists of four major DTs: DT for diagnosis (DT-D), DT for prognosis (DT-P), DT for strategy inventory (DT-SI), and DT for strategy assessment (DT-SA). As a preliminary study of NAMAC trustworthiness, this work only considers DT-D, DT-P, the discrepancy checker, and their interactions to narrow the research scope. Compared to the original NAMAC operational workflow, this study assumes a perfect strategy optimization process. Moreover, this study integrates DCA into DT-D for identifying interpolation and extrapolation conditions. As shown in Figure 1, this study simplifies the NAMAC operational workflow with three major steps:

1) At the recommendation time, t_{rcmd}, the DT-D will estimate unobservable safety significant factors (SSFs) to restore the complete reactor states. This step consists of four substeps to ensure the quality of DT-D estimations. A DT-D will be selected from the given DT hub. Its data coverage condition will be determined based on the UQ results. If the uncertainty satisfies the requirements, DT-P will use DT-D outputs. Meanwhile, the discrepancy checker will use the DT-D coverage condition and uncertainty. This study assumes that a DT-D satisfying all the requirements can always be found from the DT hub.

2) In the complete NAMAC operational workflow [2], DT-SI will sample a set of mitigation strategies at t_{rcmd}. DT-P will forecast the reactor future for each strategy, based on which

DT-SA will determine the optimal strategy and recommend it to operators. This study ignores the DT-SA and DT-SI uncertainties. Only the recommended mitigation strategy and corresponding DT-P predictions are considered.

3) The discrepancy checker will be activated at the ith discrepancy checking time t_{cki} to evaluate the reliability of NAMAC recommendations by comparing the expected reactor states against the actual measurements. If the discrepancy in the NAMAC recommendations does not meet the requirements, a SCRAM signal will be sent to the operator. Otherwise, the operation will be continued until the next discrepancy checking time.

2.1.2. Data coverage assessment

Neural networks (NNs) have been used to develop the NAMAC DTs [1, 2, 22, 23]. However, NN-based DTs may fail to predict unseen data in extrapolated conditions. A DCA framework is developed in [20] to detect extrapolated conditions by quantifying the epistemic uncertainty of an NN-based DT.

DCA development starts with building the knowledge base, including the training data and input and output features. Then a DT UQ will be conducted, based on which coverage thresholds will be selected. If the DCA performance meets the requirements, it will be deployed to NAMAC. Otherwise, the knowledge base, UQ algorithms, and coverage thresholds will be refined. During NAMAC operation, DT epistemic uncertainties can be quantified and compared to the refined thresholds. Extrapolated conditions can be identified if the epistemic uncertainties exceed the thresholds.

2.1.3. Discrepancy checker

Lin et al. [2] proposed a rule-based discrepancy checker comparing the discrepancies between the real reactor conditions and the expected transients to the given limits. In this study, we updated it with DCA. Figure 2 depicts the discrepancy checker workflow used in this study. Compared to the original scheme proposed in [2], our changes are:

- As discussed in Section 2.1.1, at the recommendation time t_{rcmd}, the data coverage of DTDs selected from DT hubs will be checked. The updated discrepancy checker will use this

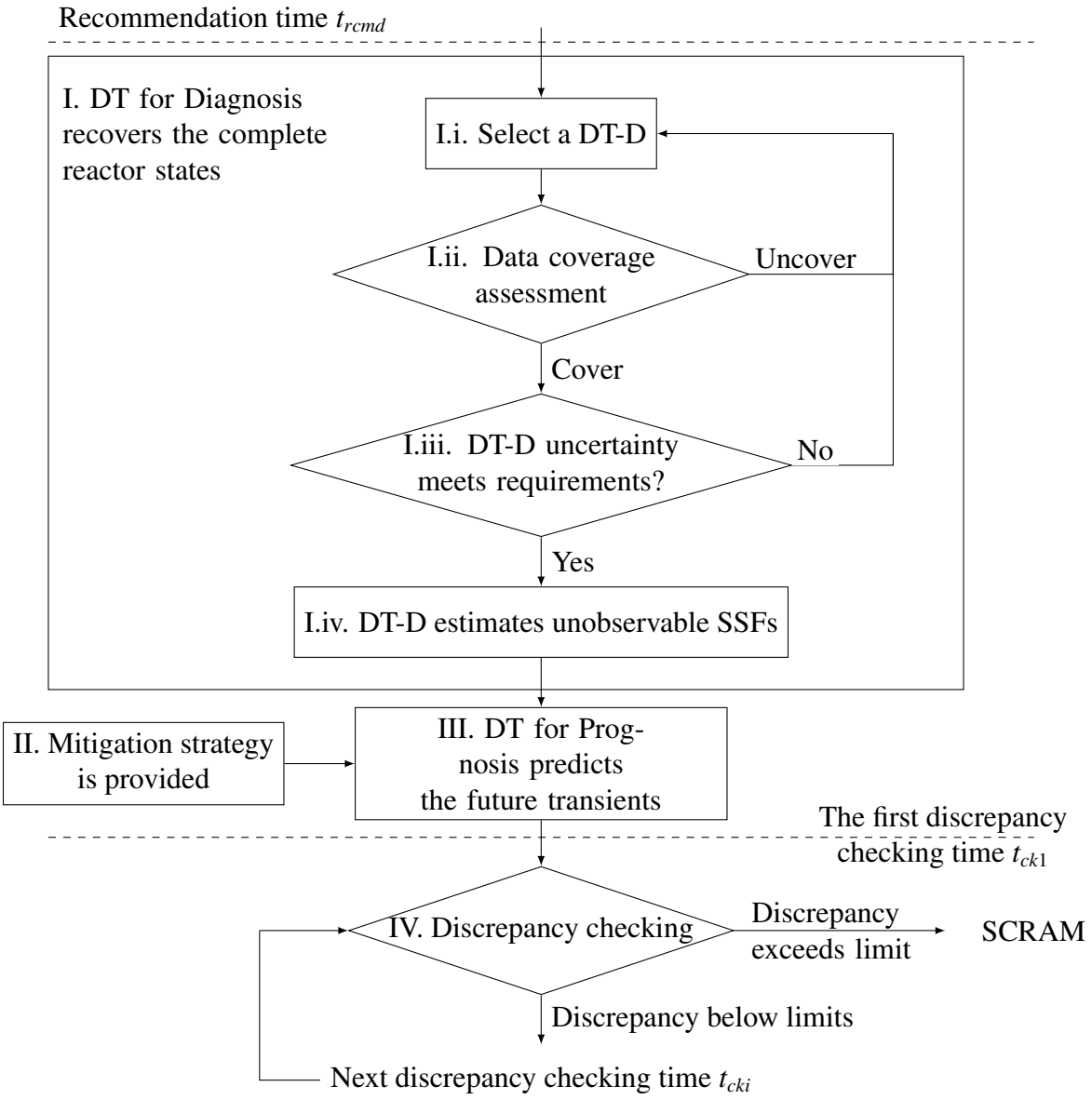

Figure 1: Simplified NAMAC operational workflow

information to evaluate NAMAC trustworthiness.

- At the discrepancy checking time $t_{ck} = \{t_{ck1}, \ldots, t_{cki}, \ldots, t_{ckN}\}$, the data coverage condition of the DT-D inferring T_{PFCL} will be assessed. If the DT-D does not cover the reactor states at t_{ck}, the discrepancy checker will assume that the NAMAC discrepancy exceeds the limit and will generate a SCRAM command to shut down the reactor.

- At t_{ck}, the local discrepancies are used as discrepancy metrics ζ. Instead of measuring the time-averaged discrepancy as proposed in [2], the discrepancy checker in this study takes a snapshot of discrepancies at t_{ck} and compares them to the discrepancy limits ξ.

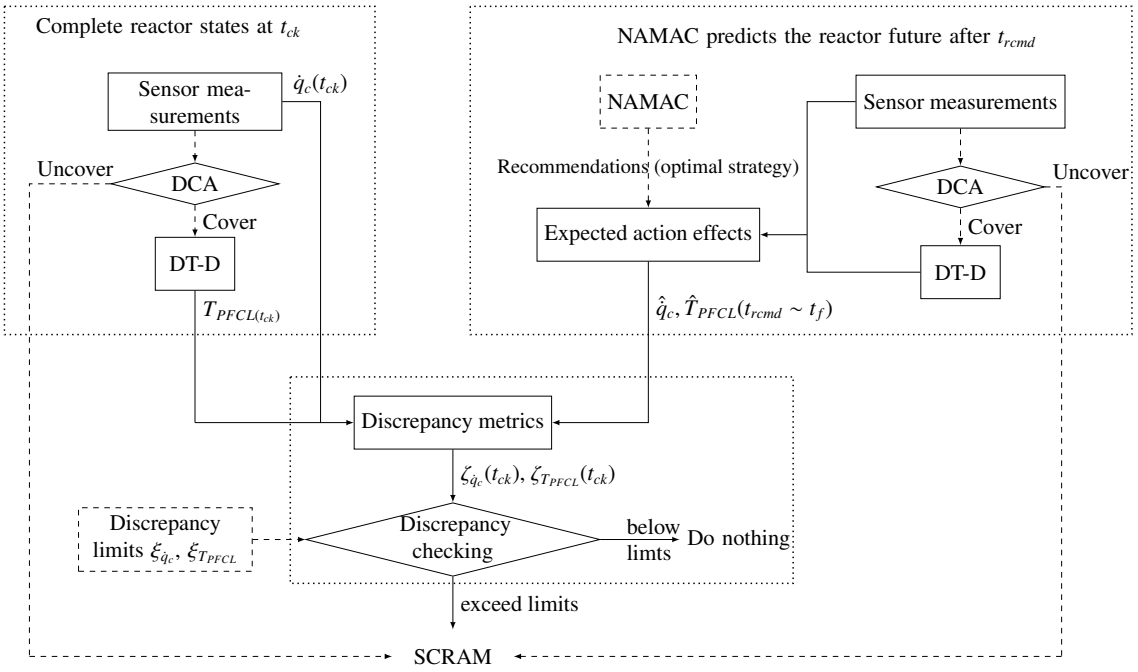

Figure 2: Discrepancy checking process with DCA updated from [2]

2.1.4. Digital-twin development and assessment process

Lin et al. [24, 25] proposed a digital-twin development and assessment process (DT-DAP) based on the Evaluation Model Development and Assessment Process [26], which can improve the reliability and robustness of DTs used in the NAMAC system. The goal of DT-DAP is to ensure that the DT development, assessment, and decision-making process are transparent, consistent, and able to fulfill the expectations established in the planning stage [25].

As shown in Figure B.17, the DT-DAP consists of five elements [24, 25]:

- Element 1 establishes requirements for DTs and the NAMAC system, such as the system purposes, the quantities of interest (QoIs), and the systems to be developed.

- Element 2 develops the knowledge base. Training and testing data will be collected or generated by simulation tools following the defined issue space and stored in a data repository. Other required information, including safety design, will also be collected and stored.
- DTs are developed in Element 3 based on the established requirements and constructed knowledge base. A training plan will be generated, and different DTs will be connected.
- In Element 4, the DT is assessed. The bottom-up approach focuses on the uncertainties of separate DTs, and the top-down approach deals with the overall quality of the integrated system.
- Finally, Element 5 determines whether the developed DTs and NAMAC system are adequate.

2.2. Predictive capability maturity model

Oberkampf et al. [27] developed the Predictive Capability Maturity Model (PCMM) to assess the maturity level of modeling and simulation (M&S) efforts. As shown in Table 1, the original PCMM matrix involves six elements [27]. Their maturity levels are qualitatively assessed, and the target maturity levels are application oriented [28]. High maturity levels are required for high-consequence applications and applications where M&S has significant impacts. Recently, PCMM has been adopted to assess the adequacy of M&S efforts [28, 29].

Table 1: PCMM table (derived from [27, 28])

Attribute	Maturity			
	Level 0	Level 1	Level 2	Level 3
Representation and Geometric Fidelity	Low $\overset{\text{Maturity}}{\Longrightarrow}$ High			
Physics and Material Model Fidelity				
Code Verification	Low $\overset{\text{Consequence}}{\Longrightarrow}$ High			
Solution Verification				
Model Validation	Low $\overset{\text{M\&S impacts}}{\Longrightarrow}$ High			
Uncertainty Quantification and Sensitivity Analysis				

3. Methodology

3.1. Trustworthiness modeling and evaluation workflow

This study aims to demonstrate how NAMAC trustworthiness can be modeled and evaluated. Figure 3 depicts the scheme used in this work. The process steps are:

- First, NAMAC trustworthiness will be conceptually modeled using trustworthiness ontologies [6, 7, 10, 11]. A rudimentary trustworthiness ontology will serve as the basis for both development-time and run-time trustworthiness modeling. Then, based on the NAMAC development and operation, detailed development-time and run-time trustworthiness ontologies will be constructed separately.

- Second, the conceptual NAMAC trustworthiness will be converted into numerical models to evaluate NAMAC trustworthiness. The trustworthiness ontologies will be converted into Bayesian networks (BNs) [30] to support the quantitative trustworthiness evaluation.

- Third, based on the trustworthiness model, required evidence will be collected and characterized. For development-time trustworthiness, evidence will be collected from development and assessment works on NAMAC, which is further formalized based on PCMM [27] and DT-DAP [24, 25]. Evidence will be collected from DCA and discrepancy checking results during the NAMAC operation for run-time trustworthiness.

- Fourth, evidence will be integrated through the numerical trustworthiness model to quantitatively evaluate NAMAC trustworthiness. Transferring qualitative evidence to some quantitative scales may be required.

3.2. Ontology-based trustworthiness modeling

In this study, we adopt the perspective that trustworthiness is a multi-attribute concept [3, 6–11] and hierarchically relate the NAMAC trustworthiness concepts following [3, 5–7], as depicted in Figure 4. NAMAC trustworthiness is treated as a collection of numerous trustworthiness concerns, and those concerns may be further decomposed into subconcerns. The so-called trustworthiness

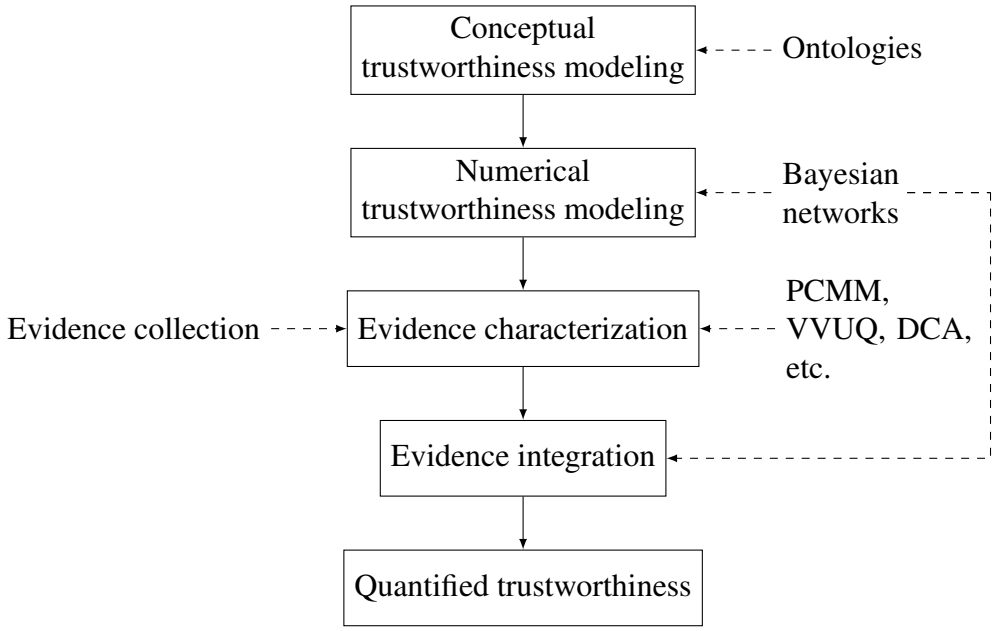

Figure 3: Scheme for NAMAC trustworthiness modeling and evaluation

properties, which are properties, features, and characteristics of the NAMAC system, can then address the trustworthiness.

From trustworthiness to trustworthiness properties, the dependence of the trustworthiness concept on NAMAC becomes stronger [3]. The top of Figure 4 is trustworthiness, a general concern for most systems and the most abstract concept. Then, trustworthiness will be supported by the less abstract concerns and subconcerns. NAMAC properties, including DTs, operational workflow, and development and assessment processes, address those concerns. Compared to concerns in the top node, elements in lower hierarchies have richer technical details but lower levels of abstraction. The bottom level, denoted as trustworthiness properties, is directly related to NAMAC development and evaluation results.

Based on the hierarchical trustworthiness model shown in Figure 4, we utilized the ontology-based method developed in [7, 11] to conceptualize and visualize NAMAC trustworthiness. An ontology can support the hierarchical and logical representation of an entity [6, 10], and the ontology-based methods have been utilized to logically conceptualize the multiattribute trustworthiness [6, 7, 10, 11, 31]. It can conceptually model the system trustworthiness and the relations between different trustworthiness concepts. An ontology-based method can enhance the trans-

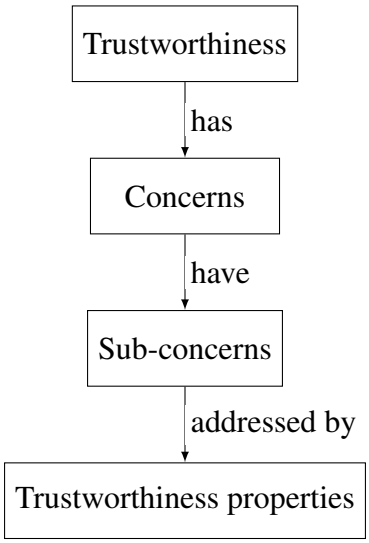

Figure 4: Concepts of hierarchical modeling of the NAMAC trustworthiness (derived based on [3, 6, 7])

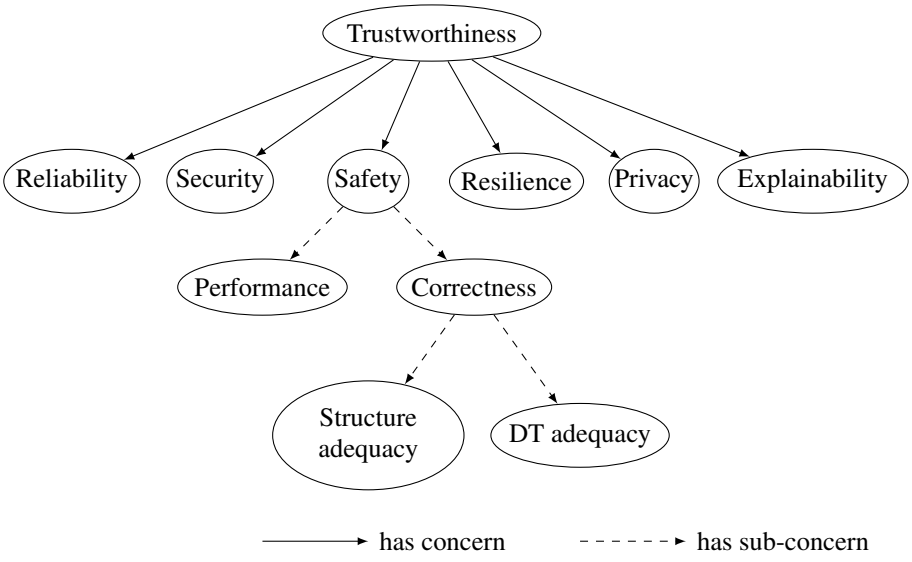

Figure 5: Rudimentary trustworthiness ontology for the NAMAC system

parency of the trustworthiness modeling and evaluation processes, since it explicitly and visually connect different trustworthiness concepts.

Figure 5 depicts the rudimentary NAMAC trustworthiness ontology from trustworthiness to selected high-level trustworthiness concerns. In the NAMAC trustworthiness ontologies, trustworthiness concerns and properties are represented by individuals, and the relations between them are characterized by the edges linking ontology individuals. The NIST CPSF [5] identifies five

concerns for trustworthiness: reliability, safety, security, resilience, and privacy, which are also frequently mentioned in different literature sources [3, 8–10, 31, 32]. We use them in the NAMAC trustworthiness ontology as high-level concerns. In addition, the explainability of AI methods and ML models is becoming an increasingly important problem [33, 34]. With the extensive use of ML models in the NAMAC system, explainability is added to the ontology as the sixth high-level concern.

Although six high-level concerns are identified for NAMAC trustworthiness, this preliminary study only focuses on safety, which may be the most significant concern of high-consequence safety-critical nuclear reactor systems. The rest of this study will omit other concerns from the ontologies. Based on the definitions of safety provided in [3, 10, 31], we define NAMAC safety as *the operation of NAMAC will not cause damage, disastrous consequences, or harm to the users and the environment*.

NAMAC safety stems from whether NAMAC recommendations will lead to safety concerns on nuclear reactors and, ultimately, the environment and public health. An incorrect or late recommendation could lead to a reactor trip and component and system failures during accident scenarios. Accordingly, the safety concern has two subconcerns, correctness and performance, based on the trustworthiness attributes reviewed in [3]. Correctness refers to whether the NAMAC system can bring the reactor to the desired conditions by providing proper recommendations to reactor operators. Performance refers to how long it takes NAMAC to make recommendations. Since NAMAC recommendations are derived based on DTs integrated by the NAMAC operational workflow, correctness can be further branched into two subconcerns: structure adequacy and DT adequacy. Similar to other NAMAC research [1], this study assumes that NAMAC can always provide on-time recommendations, so correctness is the only focus. Overall, this study simplifies the trustworthiness ontology models and assumes that NAMAC trustworthiness only depends on the correctness of NAMAC recommendations.

3.3. Development-time and run-time NAMAC trustworthiness

As discussed in [3], trustworthiness is dynamic and should be continuously re-evaluated and maintained at run-time. The NAMAC system consists of three layers [1]: knowledge base, de-

velopmental layer, and operational layer. Different processes and assumptions are involved in the developmental and operational layers. DTs are trained and assessed in the developmental layer, while they will not be tuned during operation. Moreover, DT training in the developmental layer may be designed to accommodate training data and model uncertainty. However, the real-time operation data may be subject to significant uncertainty exceeding the acceptable limits. Therefore, requirements for a trustworthy NAMAC in the developmental and operational layers may differ. This study distinguishes development-time trustworthiness from run-time trustworthiness and separately models and evaluates them.

Although development-time and run-time trustworthiness share the same high-level concerns identified in the rudimentary ontology, the low-level subconcerns and trustworthiness properties distinctly differ. In addition, different philosophies are followed when transferring ontologies to BNs. The development-time trustworthiness referenced the DT-DAP, Evaluation Model Development and Assessment Process, and NAMAC development practices, while the run-time trustworthiness mainly relies on the discrepancy checker and DCA. Therefore, we will discuss the development-time and run-time trustworthiness ontologies together with the evaluation results in Section 4 and Section 5, respectively.

With the developed trustworthiness ontology, this study focuses on:

- The simplified NAMAC system, as discussed in Section 2.1.1
- The development-time NAMAC trustworthiness and run-time trustworthiness, which will be discussed in Section 4 and Section 5, respectively
- The concerns and corresponding subconcerns of correctness, as discussed in Section 3.2
- Evidence from existing works on DTs and NAMAC development and assessment, which will be summarized in Section 4.2.3 and Section 5.3

4. Development-time Trustworthiness Evaluation

4.1. Development-time trustworthiness ontology

The NAMAC operational workflow defines the structure of the NAMAC system. This study simplifies the NAMAC operational workflow and assumes it can fulfill the basic requirements of NAMAC correctness. To evaluate the development-time NAMAC trustworthiness based on DT adequacy, we identified four subconcerns for it based on the DT-DAP described in Section 2.1.4: requirement appropriateness, data adequacy, DT training adequacy, and DT assessment adequacy. Each subconcern has corresponding trustworthiness properties identified based on existing studies of the NAMAC system and relevant works [1–3, 23, 27, 35, 36]. These attributes contribute to NAMAC trustworthiness through DT adequacy, correctness, and safety. Figure 6 shows the complete trustworthiness ontology with DT adequacy subconcerns and corresponding trustworthiness properties.

4.1.1. Requirement appropriateness

Requirement appropriateness stems from Element 1 of DT-DAP and focuses on whether proper requirements are established for DTs. Appropriate QoIs and functions should be identified for DTs such that they can be used to fulfill the requirements of the NAMAC system. The current development of the NAMAC system does not explicitly focus on whether the requirements are proper. As suggested in [3], we identified requirements engineering [37] as a potential trustworthiness property addressing requirement appropriateness. Requirements engineering has yet to be explicitly implemented to develop the NAMAC system, so we will not discuss it here.

4.1.2. Data adequacy

In Element 2 of DT-DAP, existing data will be collected and new data will be generated to construct the knowledge base. ML models are widely used in the NAMAC system to develop DTs, and the DT performance is significantly impacted by the adequacy of training data. Therefore, whether data used to develop DTs are adequate to support the functions of DTs and the entire NAMAC system may be of concern.

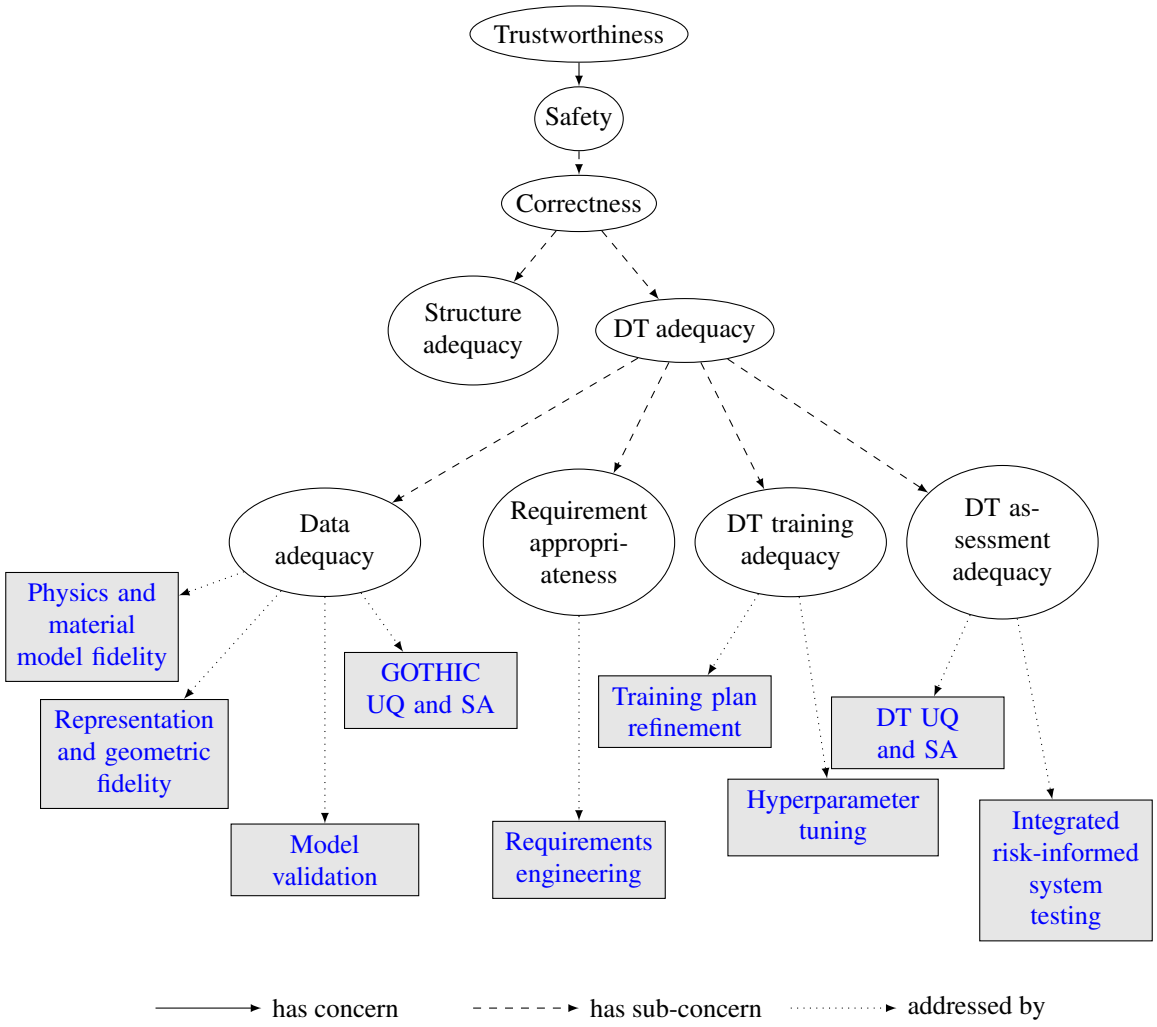

Figure 6: Trustworthiness ontology used for the development-time trustworthiness evaluation

The GOTHIC EBR-II model [38] generates the data currently used to develop the NAMAC system and DTs, so data adequacy in this study emphasizes whether the simulation results can capture the behaviors of real reactors. We examined data adequacy based on the PCMM described in Section 2.2. The original PCMM matrix consists of six attributes [27]. This study assumed that the GOTHIC code is well verified, so the four properties in Table 3 are adapted from the rest of the four attributes of the PCMM matrix to address data adequacy.

4.1.3. DT training adequacy

The concern of DT training adequacy stems from Element 3 of DT-DAP. Based on the knowledge base and requirements for DTs, NN-based DTs will be trained. Whether DTs are sufficiently trained can impact DT adequacy. Based on the existing works developing NN-based NAMAC DTs, this work identifies two trustworthiness properties addressing DT training adequacy, as listed in Table 4.

4.1.4. DT assessment adequacy

DT assessment adequacy stems from Element 4 of DT-DAP, which assesses NAMAC DTs through bottom-up and top-down approaches. It will directly affect the credibility of the assessed DT adequacy. In [25], the UQ of separate DTs is suggested for a bottom-up assessment, and integrated risk-informed analysis is identified as a critical part of the top-down approach. Following this, this study identifies two properties addressing DT assessment adequacy, as summarized in Table 5.

4.2. Development-time trustworthiness evaluation using Bayesian network

The updated development-time trustworthiness ontology conceptually models the NAMAC trustworthiness during development. In [7], the trustworthiness ontology is directly used to reason CPS trustworthiness and the individuals in the ontology only have two possible states: satisfied or not satisfied. A concern will be satisfied if all the trustworthiness properties directly or indirectly addressing it through subconcerns are satisfied. However, it may be difficult to fully satisfy the trustworthiness concern of a complex nuclear engineering system. Therefore, this study evaluates NAMAC trustworthiness based on the degree to which trustworthiness concerns and properties are satisfied. The satisfaction degree of a concern will be the weighted average of satisfaction degrees of its subconcerns or properties.

Athe and Dinh [28] and Lin and Dinh [29] use a BN for a multiattribute maturity and adequacy assessment. This study adopts a similar methodology to evaluate development-time NAMAC trustworthiness by converting the trustworthiness ontology to a BN shown in Figure 7. The nodes in a BN represent the claims that concerns or properties are satisfied, and the directed edges will

characterize the relations between concerns and properties. The top node of the BN represents NAMAC correctness, and root nodes will characterize trustworthiness properties. The BN nodes will have two possible states: True or False. The degree to which a concern or property is satisfied will be characterized by the evidence-backed belief that the corresponding node state is True.

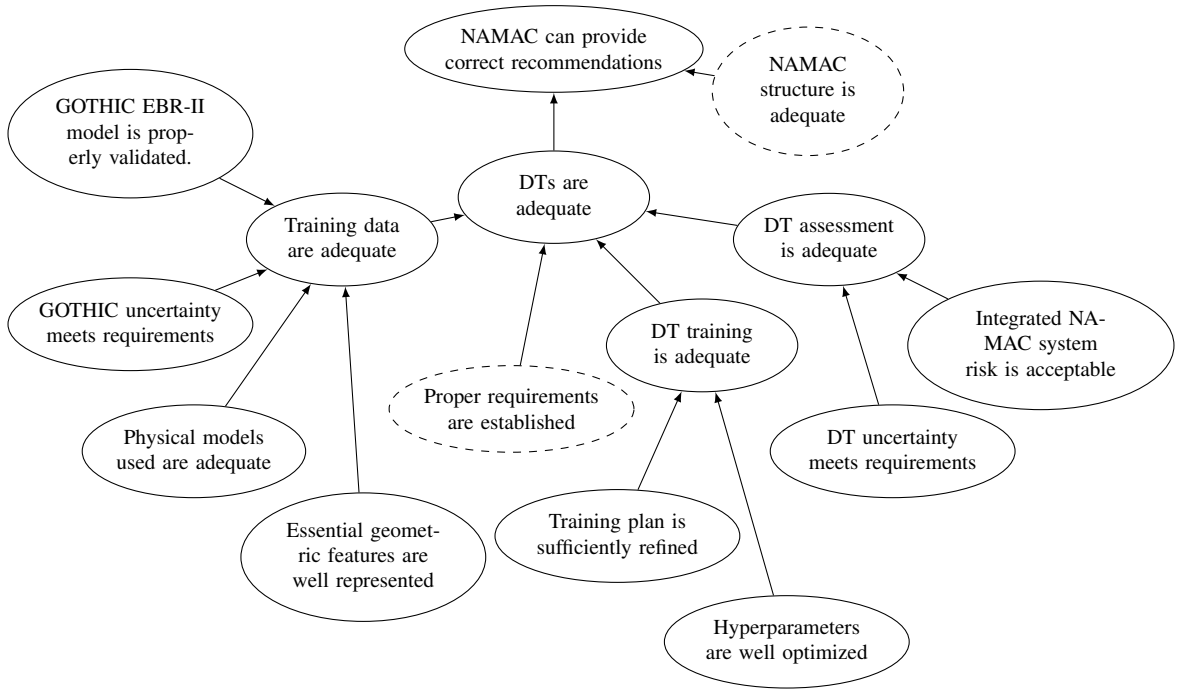

Figure 7: BN for the development-time trustworthiness evaluation (states of the dashed nodes are assumed to be True with full confidence)

After constructing the BN, the conditional probability distributions and root node beliefs should be provided to finish the evaluation. This study uses conditional probability tables (CPTs) to summarize the conditional probability distributions. The root node beliefs will be determined by collecting and characterzing evidence.

4.2.1. Determination of root node beliefs

The degrees to which a trustworthiness property is satisfied will be characterized by $P(N_{rt} = \text{True})$, where N_{rt} is the corresponding root node. This study measures how well a trustworthiness property is satisfied using four discrete satisfaction levels [28], {High, Medium, Low, N/A}, considering the similarity between the satisfaction of trustworthiness properties and predictive capability maturity. This study manually assigns the satisfaction levels of trustworthiness properties

following these rules:

- *High*: The collected evidence can support the satisfaction of all essential requirements
- *Medium*: The collected evidence can only support the satisfaction of some requirements
- *Low*: Only few requirements are satisfied based on the supporting evidence collected
- *N/A*: No available evidence can be collected to support the satisfaction of trustworthiness property.

Since the nodes in Figure 7 only have two possible states, the qualitative satisfaction levels of trustworthiness properties should be related to the quantitative beliefs in the states of the corresponding root nodes. Table 2 summarizes the beliefs in the root nodes given the satisfaction levels of the corresponding trustworthiness properties. For instance, if the satisfaction level of training plan refinement is Medium, the belief that the training plan is sufficiently refined will be 70 %. In this study, those values are manually assigned based on the authors' judgment.

Table 2: Beliefs in root node states based on the corresponding satisfaction levels

Satisfaction level	True	False
High	100 %	0 %
Medium	70 %	30 %
Low	20 %	80 %
N/A	0 %	100 %

4.2.2. Construction of CPTs

In Figure 7, trustworthiness concerns and properties are connected by directed edges, and CPTs characterize their relations. As mentioned, we assume that the degree to which a trustworthiness concern is satisfied will equal the weighted-averaged degree to which the trustworthiness concerns or properties directly related to it are satisfied, respectively. Therefore, CPTs will characterize the weights of different trustworthiness properties and subconcerns contributing to their children. For a child node C with J parent nodes $\{N_1, \ldots, N_j, \ldots, N_J\}$, the conditional probability can be

calculated as [28]:

$$P(C^D|N_1^{M_1},\ldots,N_j^{M_j},\ldots,N_J^{M_J}) = 1 - \sum_{\substack{j=1 \\ D\neq M_j}}^{J} w_j \tag{1}$$

where D is the state of node C, M_j is the state of node N_j, and w_j is the weight of node N_j. In this study, the possible states of a node are:

$$D \in \{True, False\} \tag{2}$$

$$M_j \in \{True, False\} \tag{3}$$

The weight w_j will satisfy the relation:

$$\sum_{j=1}^{J} w_j = 1 \tag{4}$$

4.2.3. Evidence characterization and integration

The GOTHIC EBR-II model has been benchmarked against the Shutdown Heat Removal Test (SHRT) results [38], and its behaviors are investigated when training data for NAMAC development are generated. When the NAMAC system and DTs were developed, efforts were made to improve their trustworthiness [1, 2, 22, 23]. The evidence for the development-time NAMAC trustworthiness evaluation is collected from these existing works. Table 3, Table 4, and Table 5 summarize the characterized evidence and corresponding satisfaction levels.

- *Representation and geometric fidelity*: **Medium**

  This property assesses whether sufficient geometric features are considered. It will analyze the geometric details neglected for simplification and evaluate their impacts on QoIs. Whether the nodalization is adequate to model the critical system characters will also be investigated.

  Although the computational EBR-II model ignores many geometric features, it is sufficient to capture most of the essential behaviors during the loss of flow tests [38]. However, detailed representations of fuels may be required to obtain better estimations of T_{PFCL}, which

is an essential SSF used by the developed NAMAC system. Therefore, the satisfaction level for the property of representation and geometric fidelity is considered Medium.

- *Physics and material model fidelity*: **Medium**

  This property aims to address whether the physical models used by the GOTHIC EBR-II model are adequate. It will analyze whether the models are physics-based or empirical, how well models are calibrated, and whether the calibration results apply to the condition of interest. In addition, it assesses the adequacy of multi-physics coupling.

  Although the GOTHIC code has been improved to support simulation of EBR-II reactor, including fluid property tables constructed for sodium and other molten salts and updated correlations specifically for liquid metal [38], whether the empirical models can be extrapolated to simulate the accident transients is of concern. Therefore, it is assumed that the satisfaction level of physics and material model fidelity is Medium.

- *Model validation*: **Low**

  Model validation assesses the accuracy of the GOTHIC model by comparing the simulation results with experimental results. Experimental uncertainty, data applicability, and scaling issues should also be addressed. The GOTHIC model may be validated against separate and integral effects tests.

  The GOTHIC EBR-II model is benchmarked against the SHRT experiments [38]. However, not all the variables used for DT training are measured. One critical parameter, T_{PFCL}, is not compared to experimental results. In addition, the SHRT experiments do not fully cover the conditions that the developed NAMAC system considers. Therefore, the satisfaction level of model validation is set to Low in this study.

- *GOTHIC UQ and sensitivity analysis (SA)*: **Low**

  The GOTHIC UQ and SA address how thoroughly and comprehensively the uncertainties and sensitivities are quantified. Whether all possible uncertainty sources are identified and whether the uncertainties are appropriately characterized and propagated through the GOTHIC model will be investigated.

UQ of the EBR-II model used to generate the NAMAC data has not been comprehensively investigated. The GOTHIC data are currently assumed uncertainty-free. Although generating a GOTHIC database with different manipulated variables can be considered as SA efforts, major gaps exist for the UQ and SA of GOTHIC, and the satisfaction level is set to Low.

- *Training plan refinement*: **Medium**

  Training plan refinement selects the proper ML model, input features, and training data to fulfill the DT requirements.

  As part of the refinement, different ML algorithms are compared [35], and optimal input features are selected [36]. However, the input features are selected based on the model performance, and the physical interactions between input features are not considered. Comprehensive feature engineering may be required to optimize the input features of DTs. In addition, training data refinement is suggested to balance the model performance and data generation cost. Therefore, the satisfaction level of training plan refinement in this study is Medium.

- *Hyperparameter tuning*: **High**

  Hyperparameter tuning focuses on optimizing ML model hyperparameters, which include the learning rate, number of neurons, and number of layers.

  To the best of the author's knowledge, all the essential hyperparameters including learning rate, number of layers, and sequence length are considered and can be optimized using Bayesian optimization [2]. In addition, different learning rate schedulers are compared to select the optimal one [36]. Therefore, the satisfaction level of hyperparameter tuning is High.

- *DT UQ and SA*: **Low**

  DT UQ and SA focus on the UQ of separate DTs. The uncertainty sources should be identified, and their impacts should be comprehensively analyzed. The corresponding uncertainties should then be adequately quantified.

Preliminary studies of SA and DT UQ have been conducted [1, 22, 35]. However, UQ of NNs is still being actively investigated, and NN uncertainties are not consistently interpreted and utilized through literatures. In addition, some essential uncertainty sources, such as training data uncertainties, are not properly captured and their impacts are not comprehensively investigated. Therefore, major gaps exist for DT UQ and SA, and the satisfaction level is Low.

- *Integrated risk-informed system testing*: **Low**

  The integrated risk-informed system testing focuses on the overall performance of the NAMAC system integrating different DTs. DT errors and uncertainties should be comprehensively characterized and propagated through the NAMAC system. Their impacts on NAMAC risk should be appropriately analyzed.

  The current studies of DT error and uncertainty propagations are mostly qualitative and empirical. Although the necessity of integrated NAMAC risk assessment has been identified, it has not been comprehensively assessed. Therefore, we assign the satisfaction level of this property to Low.

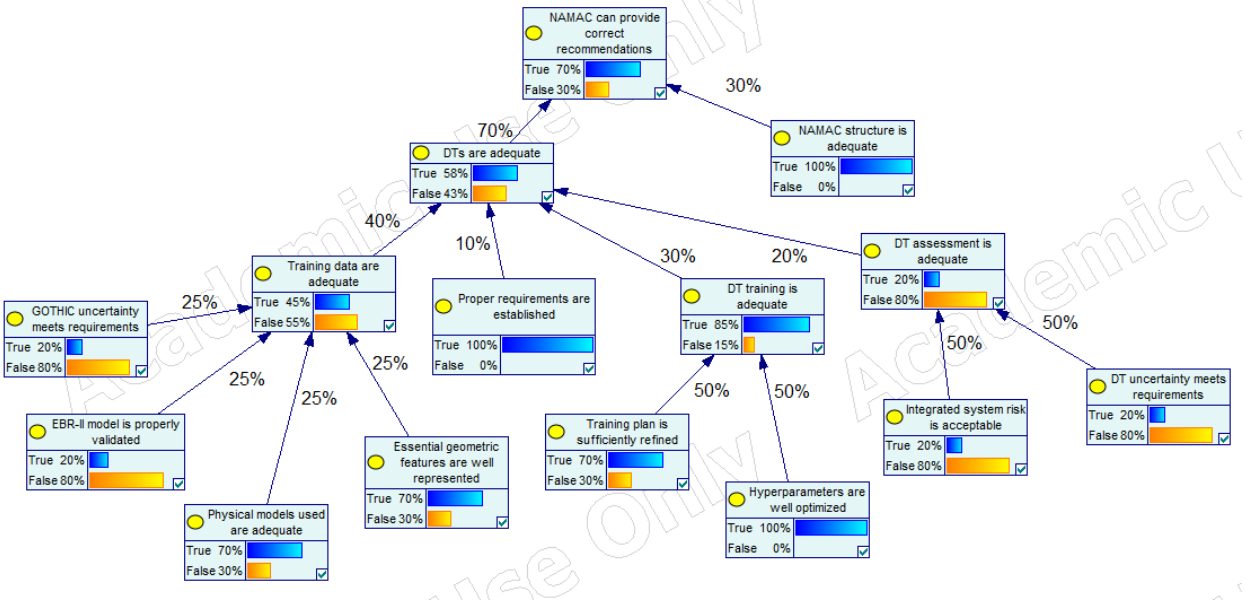

Figure 8: Evaluation of development-time NAMAC correctness based on the BN shown in Figure 7

Table 3: Characterization of evidence for data adequacy

Property	Supporting evidence and identified gaps	Satisfaction level
Representation and geometric fidelity	• GOTHIC EBR-II model is constructed based on publicly available information [38] • The computational model with simplifications can simulate essential EBR-II reactor behaviors during SHRT tests [38]	Medium
	Identified gaps: • Detailed representation of fuel may be required to obtain better estimations of T_{PFCL}	
Physics and material model fidelity	• Point reactor kinetics model is used to model the power generation [38] • Sodium property table is generated based on existing literature [38] • The interface heat mass and heat transfer model and wall friction model are improved [38]	Medium
	Identified gaps: • The empirical models should be extrapolated to accident transients with caution	
Model validation	• GOTHIC EBR-II model is benchmarked to EBR-II SHRT-17 and SHRT-45R tests [38] • The critical steady-state variables can be accurately predicted by the GOTHIC EBR-II model [38] • The general trends of the transient are captured, but some features, such as some temperature peaks, are not observable from the simulation results [38]	Low
	Identified gaps: • Some state variables, such as the peak fuel temperature, are not validated against experimental data • The validation space does not fully cover the conditions of interest	
GOTHIC UQ and SA	• Generating GOTHIC databases by varying the manipulated variables can be considered efforts of SA	Low
	Identified gaps: • The GOTHIC simulations are currently assumed to be uncertainty-free • The GOTHIC EBR-II model form and model parameter uncertainties are not characterized	

Figure 8 demonstrates the development-time NAMAC trustworthiness evaluation using the evidence characterized in Table 3, Table 4, and Table 5. The BN is constructed using the Ge-

Table 4: Characterization of evidence for DT training adequacy

Property	Supporting evidence and identified gaps	Satisfaction level
Training plan refinement	• Proper input features are suggested for the DT-D developed using recurrent neural networks (RNNs) by investigating the behaviors of models with different input features [23] • Different NN-based models are compared for time-series predictions [2, 22]	Medium
	Identified gaps: • Comprehensive feature engineering may be required to address the underlying physics between different input features	
Hyperparameter tuning	• Learning rate scheduling policy is selected from several candidates [36] • Hyperparameters are tuned with sequential model-based approach using Bayesian inference [2]	High

NIe modeler [40], and the weights of each node are labeled as percentages. We assume that the requirements for DTs are properly established and that the NAMAC structure is adequate. DT adequacy is of more concern than the NAMAC structure when NAMAC correctness is considered. Training data are vital to the success of NN-based DTs, so a higher weight is assigned to data adequacy. Based on the authors' judgment, subconcerns' importance to DT adequacy is ranked as data adequacy (40 %) > training adequacy (30 %) > assessment adequacy (20 %) > requirement appropriateness (10 %). Trustworthiness properties branched to a concern will have an equal contribution. With all these assumptions, our belief that NAMAC can provide correct recommendations is 70 %.

4.2.4. Sensitivity analysis

In the previous section, NAMAC correctness is only 70 % based on the authors' judgment. As the first step towards a more trustworthy NAMAC, we use SA to identify the properties with the most significant contribution to development-time trustworthiness.

In this study, the sensitivity is analyzed using Sobol indices [41], which will be estimated by an estimator updated in [42]. The SMILE engine [43] converts the BN to a computational model.

Table 5: Characterization of evidence for DT assessment adequacy

Property	Supporting evidence and identified gaps	Satisfaction level
DT UQ and SA	• RNN-based DT-D sensitivity to random noise of input features is assessed [23] • Impacts of inputs from failed sensors on DT-D results are studied [1] • Model parameter uncertainty and model hyperparameter uncertainty are quantified to capture the uncertainty of RNN-based DT-P [22]	Low
	Identified gaps: • The robustness of the UQ results is not assessed • Some uncertainties, such as the training data uncertainty, are not properly addressed	
Integrated risk-informed system testing	• Impacts of error propagation through DTs are qualitatively assessed [1] • How DT-D uncertainty propagate through DT-P is investigated for a simplified NAMAC [39]	Low
	Identified gaps: • NAMAC risk is not comprehensively assessed • Characterization and propagation of DT uncertainties are qualitative and empirical/semi-empirical, which still require more robust and comprehensive investigations	

If not specified, one million sets of parameters will be randomly sampled for SA in the rest of this study.

This study does not consider that no evidence is available to characterize the satisfaction levels of trustworthiness properties. One million sets of parameters are sampled by randomly selecting the satisfaction levels from {Low, Medium, High} with equal probability. The estimated Sobol indices are summarized in Figure 9. S_i represents the impacts of the ith parameter without interactions between parameters, and S_{T_i} characterizes the total impacts of the ith parameter considering its interaction with other parameters.

As shown in Figure 9, training plan refinement and hyperparameter tuning are inherently more important than other trustworthiness properties. However, their satisfaction levels are identified as Medium and High, respectively, so development-time NAMAC trustworthiness may not significantly benefit from improving them. The other trustworthiness properties are equally important,

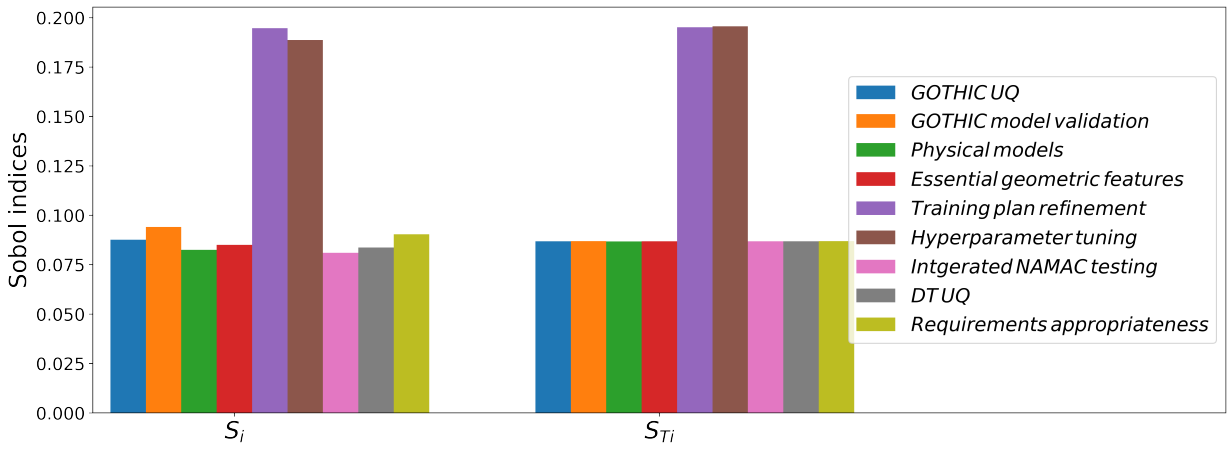

Figure 9: Estimated Sobol indices for SA of development-time trustworthiness

but some are currently identified with low satisfaction levels, such as GOTHIC model validation, DT UQ and SA, and integrated risk-informed system testing. With current assumptions, we can conclude that improving trustworthiness properties with low satisfaction levels should be the priority to make the NAMAC system more trustworthy during development.

5. Run-time Trustworthiness Re-evaluation

Although the NAMAC system may be evaluated as trustworthy during the NAMAC development process, this trustworthiness may change and should be maintained during operation. DCA and the discrepancy checker have already been developed [2, 20, 21] and can be approaches to evaluate run-time trustworthiness. This section shows how the run-time NAMAC trustworthiness can be re-evaluated using the currently developed approaches.

5.1. Run-time trustworthiness ontology

The first step is constructing the run-time trustworthiness ontology. Figure 10 depicts the run-time trustworthiness ontology updated from the ontology shown in Figure 5. Compared to the development-time trustworthiness ontology, the run-time trustworthiness ontology removes requirement appropriateness from the subconcerns of DT adequacy since the requirements for DTs are not considered dynamic during operation. In addition, DTs are not trained during NAMAC operation, so the concerns regarding DT models are lumped into model adequacy. As a result,

the run-time ontology decomposes DT adequacy into two subconcerns: model adequacy and data adequacy.

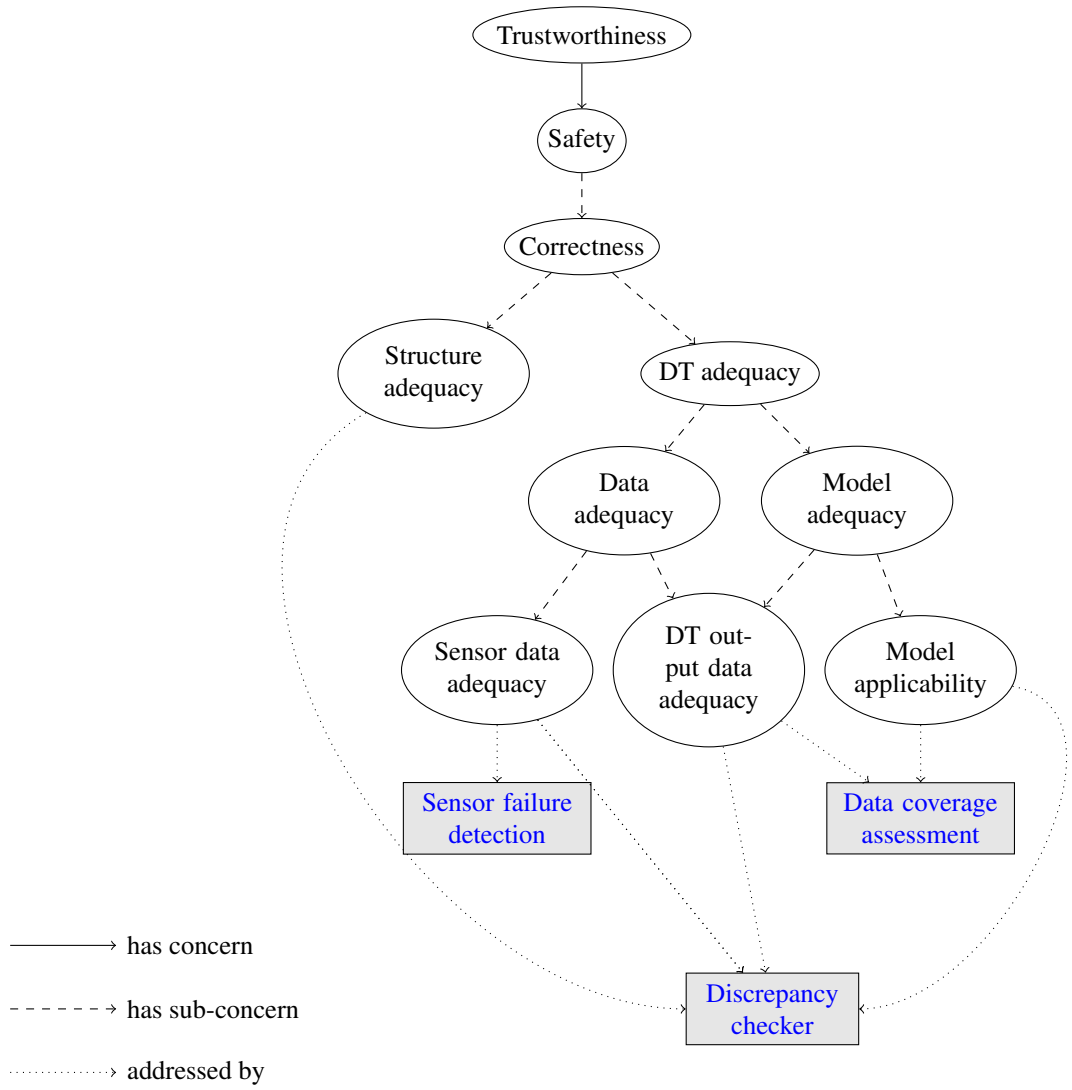

Figure 10: Trustworthiness ontology used for the run-time NAMAC trustworthiness re-evaluation

- **Data adequacy**

  Data adequacy focuses on whether data used during NAMAC operation are adequate. During operation, the data used by a DT can be externally provided by sensors and internally provided by other DTs. Therefore, data adequacy on the ontology is branched with two subconcerns: sensor data adequacy and DT output data adequacy.

Sensor measurements are usually inherently noisy. Although sensor uncertainties may be considered when DTs are developed, and the trained DT may be able to accommodate small input uncertainties, excessive sensor uncertainties can negatively impact the DT performance and lead to incorrect NAMAC recommendations.

In addition, although DTs may be well modeled and sufficiently assessed during NAMAC development, their outputs will not be uncertainty-free. As proposed, DTs used during NAMAC operation will be selected from DT hubs. The output adequacy of the selected DTs cannot always be guaranteed. Therefore, DT output adequacy is a concern and will be branched to data adequacy.

- **Model adequacy**

  Model adequacy considers whether DT models satisfy the requirements. First, whether the selected DT model applies to the reactor condition is a concern. By design, a single DT does not apply to all the possible reactor conditions, so the DT model applicability should be evaluated to check DT model adequacy. In addition, DT outputs are not uncertainty-free, which may be caused by DT model inadequacy. Although a DT applies to reactor conditions, its output errors related to a DT model may be unacceptable to other DTs. Therefore, model adequacy also depends on DT output adequacy.

We then identify the following three trustworthiness properties to address the run-time trustworthiness concerns of interest.

- *DCA*

  DCA can assess whether the pretrained DTs apply to the current reactor conditions based on data coverage conditions. As discussed in [39], DCA can also estimate the DT-D uncertainties, so it should also be used to evaluate whether DT output data are adequate. Therefore, DCA is identified as a trustworthiness property addressing subconcerns of DT output data adequacy and model applicability on the run-time trustworthiness ontology.

- *Discrepancy checker*

The discrepancy checker is designed to prevent NAMAC failures by checking discrepancies. It checks whether the discrepancies between the expected reactor behaviors and true reactor states exceed the given limits. As demonstrated in Section 2.1.1, the discrepancy checker considers the sensor uncertainties, DT-D uncertainties, and data coverage condition of DT-D. The structure adequacy and DT-P uncertainty may be implicitly addressed by comparing the estimated discrepancy to the given limits. Therefore, we identify that the discrepancy checker can be used to address all four subconcerns.

- *Sensor failure detection*

  Sensor failures may significantly weaken the sensor data adequacy, so sensor failure detection in this work is a trustworthiness property to address the concern of sensor data adequacy. However, it has not yet been implemented in the NAMAC system, and this study assumes sensor failures can be confidently detected.

5.2. Run-time trustworthiness re-evaluation using Bayesian network

The previous section identifies three trustworthiness properties to model the run-time NAMAC trustworthiness. For the simplified NAMAC system used in Section 2.1.1, the discrepancy checker updated in Section 2.1.3 already considers the DCA of the DT-D and sensor uncertainty. Therefore, this study treated it as an integrated component addressing the run-time NAMAC correctness. However, the discrepancy checker only provides deterministic evaluations of run-time NAMAC correctness and omits the uncertainties involved in the discrepancy-checking process. Therefore, this study constructs a BN based on the discrepancy-checking process to accommodate a probabilistic run-time trustworthiness evaluation under uncertainty.

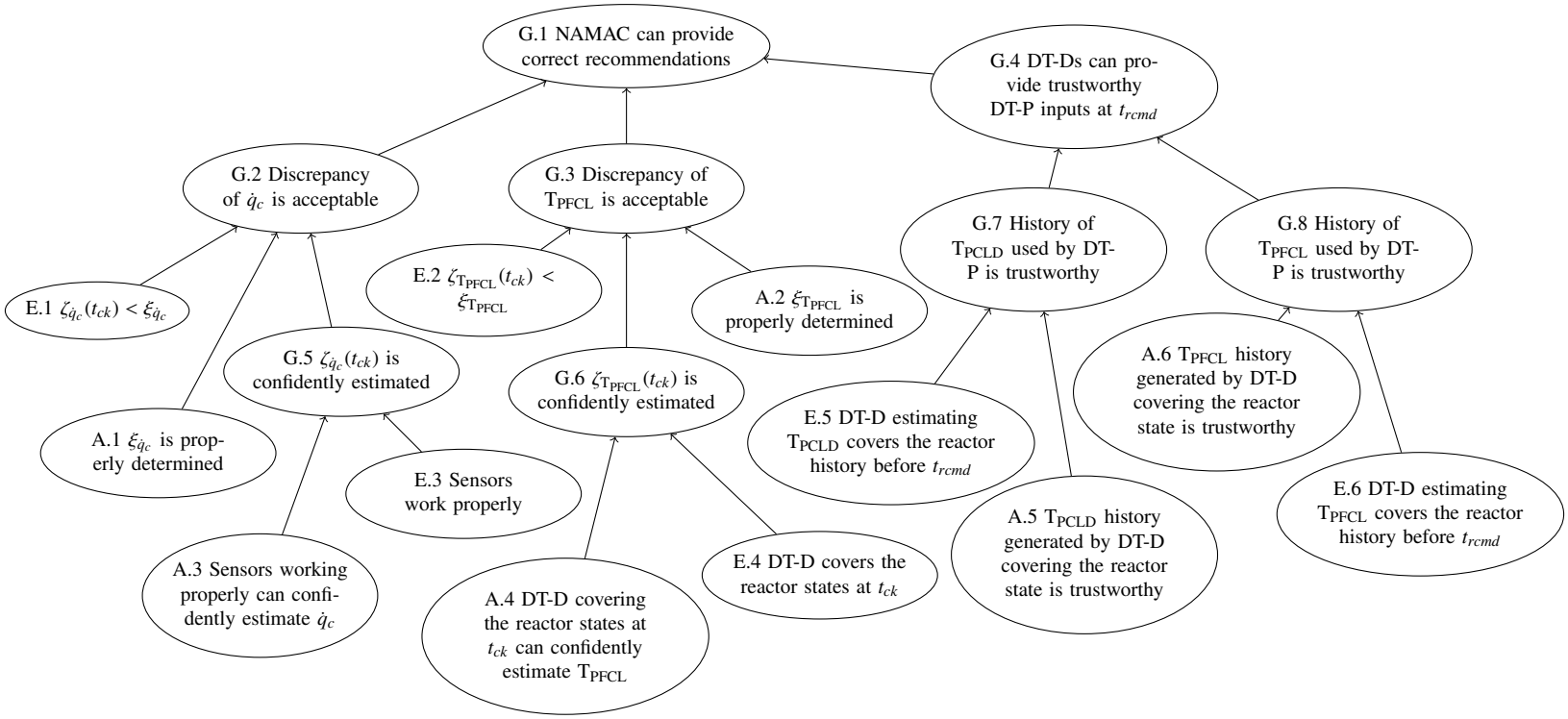

Figure 11: BN to evaluate the run-time NAMAC correctness

Figure 11 depicts the BN constructed for the run-time NAMAC trustworthiness re-evaluation. The BN focuses on NAMAC correctness and only applies to the simplified NAMAC system described in Section 2.1.1. Based on goal structuring notation [44] and work conducted in [45], this study categorizes the nodes and corresponding claims in Figure 11 into three types to simplify the discussion:

- **Goal** (G.1–G.8)

  Goals are the objectives related to the requirements of NAMAC trustworthiness. For instance, the claim that "NAMAC can provide correct recommendations" is a goal related to correctness.

- **Evidence** (E.1–E.6)

  Evidence refers to the facts or results collected to support the corresponding goals. For example, the fact that $\zeta_{\dot{q}_c}(t_{ck}) < \xi_{\dot{q}_c}$ is evidence backing the goal that "discrepancy of $\dot{q}_c$ is acceptable."

- **Axiom** (A.1–A.6)

  The connections between goals and evidence should be rationalized, and the uncertainties involved in the rationals should be properly captured to evaluate the run-time trustworthiness. There may exist two possible ways to represent rationals in a BN [45]: 1) implicitly encoding them in CPTs and 2) explicitly encoding them as separate nodes. Either way, the beliefs can be manually set. This study encodes them as individual nodes and groups them as axioms to simplify BN construction and SA. For example, the claim that "$\xi_{\dot{q}_c}$ is properly determined" is an axiom rationalizing the connections between G.2, G.5, and E.1.

Based on Figure 2, the discrepancy checker checks discrepancies of $\dot{q}_c$ and T_{PFCL}, so NAMAC correctness (G.1) will directly depend on the $\dot{q}_c$ and T_{PFCL} discrepancies (G.2, G.3), which will be backed by the discrepancy checking results (E.1, E.2). To check the credibility of the discrepancy-checking process, the credibility of the discrepancy limits (A.1, A.2) and that of discrepancy measurement (G.5, G.6) should be evaluated. This work assumes proper discrepancy limits are used,

so A.1 and A.2 are used to characterize the fidelity of the assumptions. Sensory data and DT-D results at t_{ck} are used to measure discrepancies, so sensor integrity (E.3) and DCA results at t_{ck} (E.4) are identified as evidence-backing claims G.5 and G.6. A.3 and A.4 rationalize their relations. In addition, although DCA is activated at t_{rcmd}, assuming DCA results are perfect is usually unacceptable. To capture the uncertainties caused by DCA, G.4 is appended to G.1. DT-P uses the results of two DT-Ds, which results in G.7 and G.8. Similar to G.5 and G.6, they are backed by evidence E.5 and E.6. A.5 and A.6 rationalize the connections.

5.3. Belief characterization

In the previous section, the structure of the BN used for the run-time NAMAC trustworthiness re-evaluation is determined. The conditional probabilities are determined in the following ways to evaluate the run-time NAMAC trustworthiness.

- For evidence nodes, we assume that P(E.n=True) will be either 1 or 0, where evidence is available or unavailable, respectively.
- For goal nodes, we assumed:

$$P(\text{G.n=True}|\text{All parents are True}) = 1 \tag{5}$$

and

$$P(\text{G.n=False}|\text{Any parent is False}) = 1. \tag{6}$$

That is, we assume that a goal will be achieved only if all its sub-goals are achieved, all the axioms are validated, and all the required pieces of evidence are available.

- For axiom nodes, we set P(A.n=True) = p_{An}, which characterizes the credibility of axiom A.n. In this study, the values of p_{An} were determined based on supporting data collected during the development of the NAMAC system and the authors' judgment.

Axioms A.1 and A.2 are related to whether the discrepancy limits are properly determined. The discrepancy limits used in [2] are only for demonstration purposes and are based on expert

judgment. They need to be comprehensively validated and assessed, so we have relatively low confidence in them and assume $p_{A1} = p_{A2} = 0.9$.

Axiom A.3 assesses the confidence in the sensor measurements produced by sensors working normally. This study assumes that $\dot{q}_c$ can always be confidently measured if no sensor failure occurs. Therefore, $p_{A3} = 1$.

Axiom A.4 is related to the DT-D DCA. At t_{ck}, a DT-D estimates the high-fidelity T_{PFCL}, which will be compared to the T_{PFCL} predicted by the DT-P. As discussed in [20], the positive predictive value (PPV) calculated by Equation 7 quantifies the ratio of true positives (TPs) obtained by DCA and the total positive instances predicted by DCA, so this study uses PPV as p_{A4}. In [39], the developed DCA approach is evaluated to support the development of the discrepancy checker. Based on the evaluation results, $p_{A4} \approx 0.99$.

$$p_{A4} = \text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{7}$$

Nodes A.5 and A.6 are also related to the DT-D DCA. p_{A5} and p_{A6} assess the credibility that if DT-D covers the reactor conditions shortly before t_{rcmd}, DT-D outputs will be trustworthy. At t_{rcmd}, N_{seq} sequential estimations for one unobservable SSF are required, where N_{seq} is the sequence length, so it is not proper to use PPV as p_{A5} and p_{A6}. Two possible treatments for p_{A5} and p_{A6} may exist:

1) DT-D estimations made at different time steps are assumed to be independent. For a single estimation, the confidence that the DT-D output is trustworthy can be estimated by PPV. For N_{seq} independent estimations, the probability p_{A5} that all the N_{seq} independent DT-D outputs are trustworthy can be calculated as $\text{PPV}^{N_{seq}}$.

2) DT-D estimations made at N_{seq} sequential time steps are not assumed to be independent. It can be assumed that if DT-D can completely cover the reactor states at N_{seq} sequential time steps, the probability p_{A5} that all DT-D outputs are trustworthy will be higher than if the N_{seq}

time steps are randomly selected. We may assume that:

$$\mathrm{PPV}^{N_{seq}} < p_{A5} < 1. \tag{8}$$

This study adopts the nonindependent treatment. However, analytical quantification of p_{A5} and p_{A6} may be difficult, so we formulated an evaluation case to empirically approximate p_{A5} and p_{A6}. For DT-P $N_{seq} = 5$, so five consecutive DT-D estimations will be treated as a datum. A datum will be treated as covered if all five DT-D estimations are evaluated as covered. If all the estimations are truly covered, it will be considered a successful datum. The ratio of the number of successful data to the total number of data evaluated as covered will be used as empirical estimations of p_{A5} and p_{A6} in this study. Based on the evaluation results obtained, $p_{A5} \approx 0.99$ and $p_{A6} \approx 1$.

With the conditional probabilities determined in this section, the belief in the run-time NAMAC correctness P(G.1=True) can be calculated. The BN is constructed using the GeNIe Modeler [40]. Figure 12 demonstrates the assessment of run-time NAMAC correctness, and the confidence that the NAMAC system provides correct recommendations during operation is 79 %. As illustrated in [45], this approach may only capture the lower limit of the confidence in NAMAC correctness since it omits the confidence if either of the goals or axioms is not satisfied.

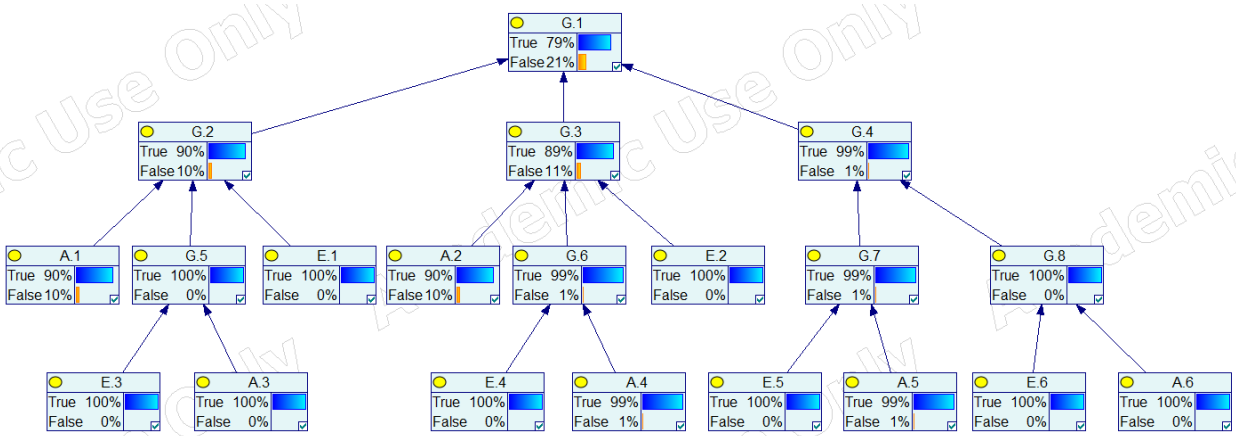

Figure 12: Evaluation of run-time NAMAC correctness based on the BN shown in Figure 11

5.4. Sensitivity analysis

This study manually assigns the axiom beliefs to evaluate the run-time NAMAC trustworthiness. We did not investigate how the credibility of p_{An} can be improved. Instead, SA is conducted to find the most influential parameters.

The belief P(G.1=True) is selected as the QoI of the sensitivity analysis. In the following discussions, P(G.1=True) will be reduced to P(G.1), and the impacts of parameters p_{An} on P(G.1) will be analyzed. Therefore, the function of interest for SA can be written as:

$$P(\text{G.1}) = f(p_{A1}, \ldots, p_{A6}). \tag{9}$$

Axiom A.3 addresses the quality of the sensors deployed in reactors, and we assume the highest confidence in it. Axioms A.4, A.5, and A.6 are based on the DT-D DCA, which has been investigated in [20], so we are more confident in them. In addition, p_{A5} and p_{A6} are slightly larger than p_{A4}. Axioms A.5 and A.6 are assumed to be of equal credibility. p_{A1} characterizes the users' understanding of the nuclear reactor and accident transients of interest, and we assumed high credibility for it in this study. The rest of the parameters quantify the confidence of the proposed approach to determine the discrepancy limit. In this study, we are least confident in those parameters. Based on those assumptions, the selected parameters are assumed to follow the distributions summarized in Table 6.

Table 6: Parameter distributions assumed for SA of run-time trustworthiness

p_{A1}	p_{A2}	p_{A3}	p_{A4}	p_{A5}	p_{A6}
$\mathcal{U}_{[0.8,1]}$	$\mathcal{U}_{[0.8,1]}$	$\mathcal{U}_{[0.98,1]}$	$\mathcal{U}_{[0.94,1]}$	$\mathcal{U}_{[0.95,1]}$	$\mathcal{U}_{[0.95,1]}$

Figure 13 depicts the estimated Sobol indices. $S_i \approxeq S_{Ti}$, so interactions between parameters are not major contributors to the total variance. p_{A1} and p_{A2} are the most influential parameters, and p_{A3} is of the least importance. The Sobol indices method is variance-based, so the importance of a parameter may be related to the variance of the parameter distributions. The rest of this section designs two experiments to investigate the impacts of the characteristics of the parameter distributions, including the mean and variance, on the estimated Sobol indices.

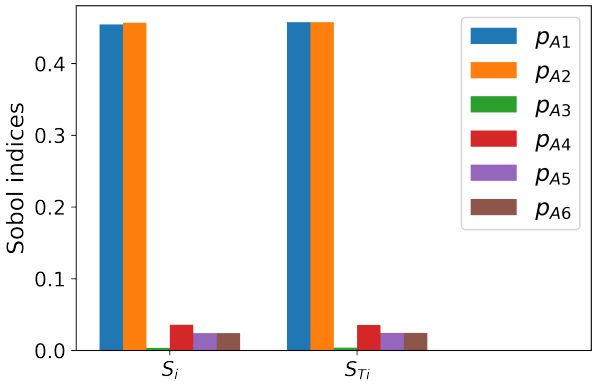

Figure 13: Estimated Sobol indices for SA of run-time trustworthiness using parameter distributions summarized in Table 6

First, Sobol indices are estimated with independent and identically distributed (i.i.d.) parameters. Figure 14 depicts the Sobol indices if the parameters are independent and identically sampled from $\mathcal{U}_{[0.8,1]}$. It shows that the selected parameters are equally important if an i.i.d. assumption is made. Then, parameters are sampled with the same variance but different mean values, as summarized in Table 7. Figure 15 shows the calculated Sobol indices. With the same variance, parameters with smaller means will be more influential.

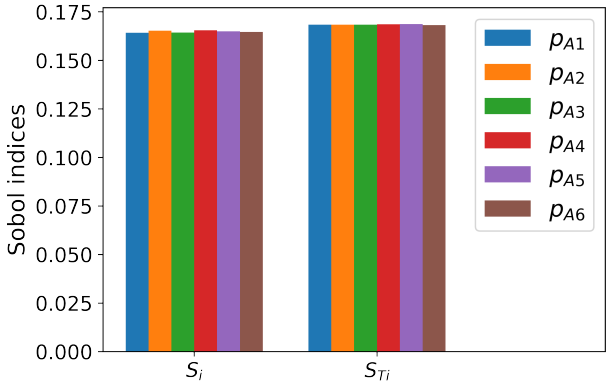

Figure 14: Estimated Sobol indices for SA of run-time trustworthiness assuming i.i.d. parameters

Table 7: Parameter distributions with the same variance assumed for SA of run-time trustworthiness

p_{A1}	p_{A2}	p_{A3}	p_{A4}	p_{A5}	p_{A6}
$\mathcal{U}_{[0.8,0.85]}$	$\mathcal{U}_{[0.8,0.85]}$	$\mathcal{U}_{[0.95,1]}$	$\mathcal{U}_{[0.92,0.97]}$	$\mathcal{U}_{[0.93,0.98]}$	$\mathcal{U}_{[0.93,0.98]}$

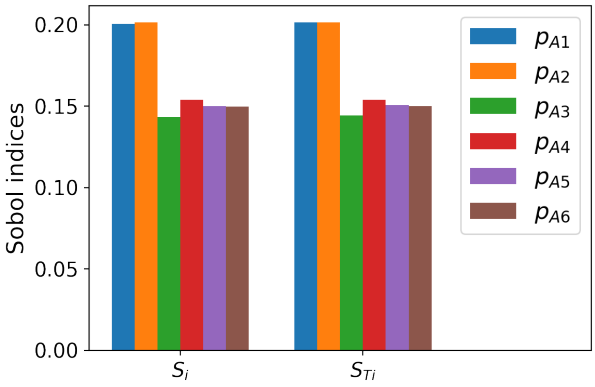

Figure 15: Estimated Sobol indices for SA of run-time trustworthiness using parameter distributions summarized in Table 7

Based on the SA results, the parameter importance mainly depends on the credibility of the corresponding axiom. An increase of the smallest p_{An} will result in the most significant increase in NAMAC correctness and trustworthiness. The credibility of the discrepancy limits used in the discrepancy checker is not assessed, and we only have low confidence in them. Therefore, more accurate discrepancy limits estimations will be the priority of future improvement.

6. Conclusion

The AI trustworthiness of simulation-based ML-enabled control systems is a critical problem in addressing the safety of modern engineering systems. This work demonstrates how the trustworthiness of a NAMAC system for advanced reactors can be modeled and evaluated. Considering the differences in key trustworthiness attributes, NAMAC trustworthiness is separately modeled and evaluated during development-time and run-time. The NAMAC trustworthiness modeling and evaluation start from trustworthiness ontologies to provide logical and hierarchical representations of the multiattribute NAMAC trustworthiness. The logical paths are then converted to BNs for quantitative trustworthiness evaluation, through which evidence is collected, characterized, and integrated.

For the simplified NAMAC, the development-time trustworthiness focuses on the adequacy of the DTs stemming from ML of EBR-II system thermal-hydraulics simulation results. It is evaluated by collecting and characterizing evidence as guided by DT-DAP and PCMM. Major

gaps exist for the validation and UQ of the GOTHIC EBR-II model, UQ of NN-based DTs, and integrated NAMAC risk analysis, which should be addressed in future work. The run-time trustworthiness is modeled and evaluated based on the DCA of DT-D and the discrepancy checker. Axioms required to close the run-time trustworthiness evaluation are identified, whose beliefs are empirically determined or assigned based on expert judgment. Sensitivity analyses suggest that no axiom is significantly more important than others, and improving the axioms with low beliefs or high uncertainties should be the focus for a trustworthy NAMAC.

Given the novelty and rapidly evolving field of AI trustworthiness, the proposed framework for trustworthiness evaluation necessarily involves subjectivity, such as determining the satisfaction levels of trustworthiness properties. The proposed approach tackles the subjectivity by enhancing the transparency of trustworthiness modeling and evaluation. With the evidence-based method, decision makers can be fully aware of the uncertainty sources and subjectivity involved in the entire process. In addition, sensitivity analyses can help prioritize and effectively allocate resources to mitigate the impact of subjectivity. Future work should also address the interactions between development-time and run-time trustworthiness. In addition, specifying metrics for trustworthiness attributes should help further reduce the assessment's subjectivity.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work is performed with the support of the ARPA-E MEITNER program under the project entitled: "Development of a Nearly Autonomous Management and Control System for Advanced Reactors" and using a GOTHIC license provided by Zachry Nuclear Engineering, Inc. GOTHIC incorporates technology developed for the electric power industry under the sponsorship of the Electric Power Research Institute. The authors would also like to thank Drs. Xu Wu, Maria Avramova, Ralph Smith, Jeffrey Lane, Paridhi Athe, and Pascal Rouxelin for their comments,

suggestions, and support. The authors appreciate Alexandria Madden for technical editing. The authors acknowledge the academic licenses of GeNIe Modeler and SMILE Engine provided by BayesFusion, LLC, the software of which is available free of charge for academic teaching and research use at http://www.bayesfusion.com/. ND acknowledges support of the U.S. Nuclear Regulatory Commission through an R&D grant "Trustworthiness of Digital-Twin-based Automation Technology in Nuclear Power Plant Operation".

Appendix A. Complete NAMAC operational workflow

Figure A.16 shows the complete NAMAC operational workflow with a discrepancy checker proposed in [2]. For readers' convenience, we here provide a brief description of it. For details of the NAMAC development, readers can refer to the references [1, 2]. For the development of DTs, references [22, 23, 46] will provide more information.

At the recommendation time t_{rcmd}, DT-D will read the observable reactor states and diagnose the complete reactor states by estimating unobservable SSFs. Then DT-SI will list all available control actions based on the complete reactor states. For each control action, DT-P will predict its impacts on reactor behaviors. With the predicted reactor states, DT-SA will rank the values of all control actions and recommend the optimal control action to operators. During the operation of the NAMAC system, a discrepancy checker will assess the NAMAC discrepancy by comparing the real-time complete reactor states inferred by DT-D and the reactor behaviors predicted by DT-P. If the discrepancy exceeds the given limit, a SCRAM command will be sent to reactor operators.

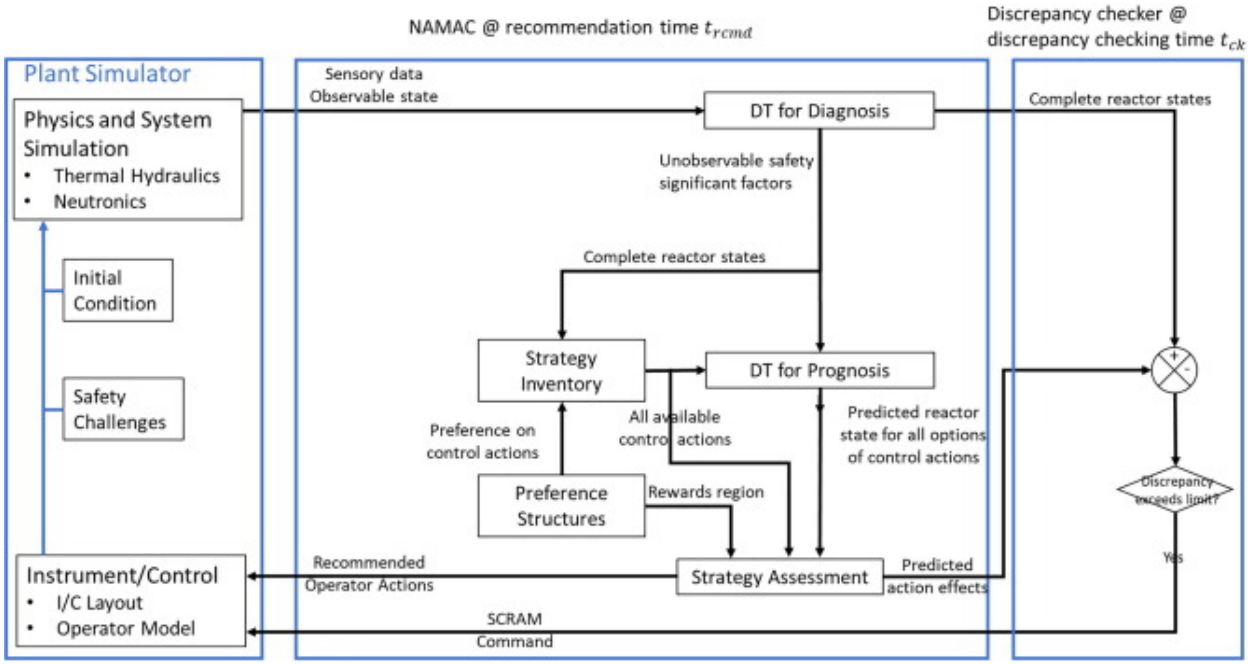

Figure A.16: Complete NAMAC operational workflow (reprinted from [2] Copyright Elsevier (2021))

Appendix B. DT-DAP scheme

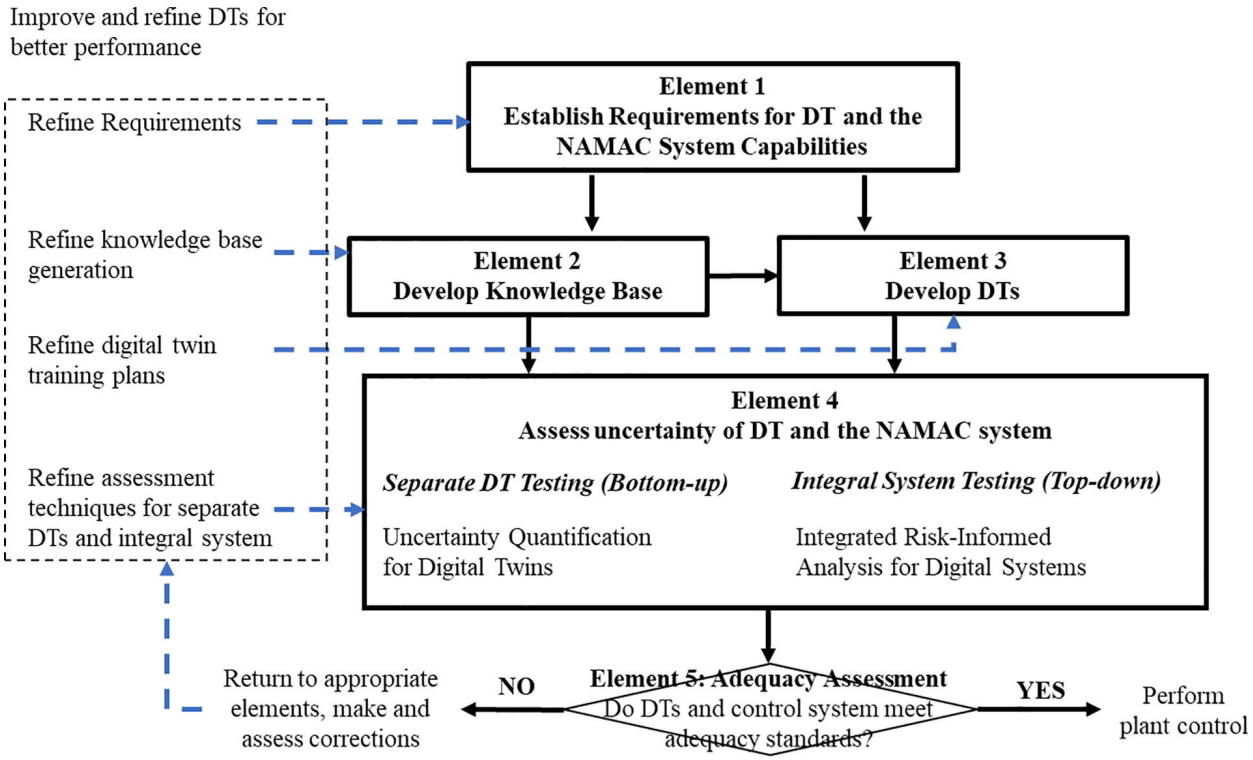

Figure B.17: DT-DAP scheme (reprinted from [25] Copyright Elsevier (2021))

References

[1] L. Lin, P. Athe, P. Rouxelin, M. Avramova, A. Gupta, R. Youngblood, J. Lane, N. Dinh, Development and assessment of a nearly autonomous management and control system for advanced reactors, Annals of Nuclear Energy 150 (2021) 107861.

[2] L. Lin, P. Athe, P. Rouxelin, M. Avramova, A. Gupta, R. Youngblood, J. Lane, N. Dinh, Digital-twin-based improvements to diagnosis, prognosis, strategy assessment, and discrepancy checking in a nearly autonomous management and control system, Annals of Nuclear Energy 166 (2022) 108715. URL: https://www.sciencedirect.com/science/article/pii/S0306454921005910. doi:https://doi.org/10.1016/j.anucene.2021.108715.

[3] N. G. Mohammadi, Trustworthy Cyber-Physical Systems: A Systematic Framework towards Design and Evaluation of Trust and Trustworthiness, 1st ed., Springer Vieweg Wiesbaden, 2019.

[4] A. BahooToroody, M. M. Abaei, O. V. Banda, P. Kujala, F. De Carlo, R. Abbassi, Prognostic health management of repairable ship systems through different autonomy degree; from current condition to fully autonomous ship, Reliability Engineering & System Safety 221 (2022) 108355. URL: https://www.sciencedirect.com/science/article/pii/S0951832022000345. doi:https://doi.org/10.1016/j.ress.2022.108355.

[5] E. Griffor, C. Greer, D. Wollman, M. Burns, Framework for Cyber-Physical Systems: Volume 1, Overview, 2017. doi:https://doi.org/10.6028/NIST.SP.1500-201.

[6] M. Balduccini, E. Griffor, M. Huth, C. Vishik, M. Burns, D. Wollman, Ontology-based reasoning about the trustworthiness of cyber-physical systems, in: Living in the Internet of Things: Cybersecurity of the IoT - 2018, 2018, pp. 1–10. doi:10.1049/cp.2018.0012.

[7] T. H. Nguyen, T. C. Son, M. Bundas, M. Balduccini, K. C. Garwood, E. R. Griffor, Reasoning about trustworthiness in cyber-physical systems using ontology-based representation and asp, in: PRIMA 2020: Principles and Practice of Multi-Agent Systems, Springer International Publishing, 2020, pp. 51–67.

[8] H. Gao, Z. Yu, X. Cong, J. Wang, Trustworthiness Evaluation of Smart Grids Using GSPN, in: 2021 IEEE International Conference on Networking, Sensing and Control (ICNSC), volume 1, IEEE, 2021, pp. 1–7.

[9] S. Ding, S.-L. Yang, C. Fu, A novel evidential reasoning based method for software trustworthiness evaluation under the uncertain and unreliable environment, Expert Systems with Applications 39 (2012) 2700–2709.

[10] J.-H. Cho, S. Xu, P. M. Hurley, M. Mackay, T. Benjamin, M. Beaumont, STRAM: Measuring the trustworthiness of computer-based systems, ACM Computing Surveys (CSUR) 51 (2019).

[11] T. H. Nguyen, M. Bundas, T. C. Son, M. Balduccini, K. C. Garwood, E. R. Griffor, Specifying and reasoning about cps through the lens of the nist cps framework, Theory and Practice of Logic Programming 23 (2023) 1029–1069. doi:10.1017/S1471068422000035.

[12] M. Dennis, T. Lalain, L. Betancourt, A. Hathaway, R. Anzalone, Artificial Intelligence Strategic Plan: Fiscal

Years 2023-2027, Technical Report NUREG-2261, United States Nuclear Regulatory Commission, Washington, DC, 2023.

[13] E. Tabassi, Artificial Intelligence Risk Management Framework (AI RMF 1.0), Technical Report, 2023. URL: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=936225. doi:https://doi.org/10.6028/NIST.AI.100-1.

[14] M. Chelouati, A. Boussif, J. Beugin, E.-M. El Koursi, Graphical safety assurance case using goal structuring notation (gsn) — challenges, opportunities and a framework for autonomous trains, Reliability Engineering & System Safety 230 (2023) 108933. URL: https://www.sciencedirect.com/science/article/pii/S0951832022005488. doi:https://doi.org/10.1016/j.ress.2022.108933.

[15] T. Zhou, T. Han, E. L. Droguett, Towards trustworthy machine fault diagnosis: A probabilistic bayesian deep learning framework, Reliability Engineering & System Safety 224 (2022) 108525. URL: https://www.sciencedirect.com/science/article/pii/S095183202200179X. doi:https://doi.org/10.1016/j.ress.2022.108525.

[16] T. Han, Y.-F. Li, Out-of-distribution detection-assisted trustworthy machinery fault diagnosis approach with uncertainty-aware deep ensembles, Reliability Engineering & System Safety 226 (2022) 108648. URL: https://www.sciencedirect.com/science/article/pii/S0951832022002836. doi:https://doi.org/10.1016/j.ress.2022.108648.

[17] T. Zhou, L. Zhang, T. Han, E. L. Droguett, A. Mosleh, F. T. Chan, An uncertainty-informed framework for trustworthy fault diagnosis in safety-critical applications, Reliability Engineering & System Safety 229 (2023) 108865. URL: https://www.sciencedirect.com/science/article/pii/S0951832022004823. doi:https://doi.org/10.1016/j.ress.2022.108865.

[18] L. Wang, H. Cao, Z. Ye, H. Xu, Bayesian large-kernel attention network for bearing remaining useful life prediction and uncertainty quantification, Reliability Engineering & System Safety 238 (2023) 109421. URL: https://www.sciencedirect.com/science/article/pii/S0951832023003356. doi:https://doi.org/10.1016/j.ress.2023.109421.

[19] J. Wang, S. Gao, L. Yu, X. Liu, F. Neri, D. Zhang, L. Kou, Uncertainty-aware trustworthy weather-driven failure risk predictor for overhead contact lines, Reliability Engineering & System Safety 242 (2024) 109734. URL: https://www.sciencedirect.com/science/article/pii/S0951832023006488. doi:https://doi.org/10.1016/j.ress.2023.109734.

[20] L. Wang, L. Lin, N. Dinh, Data coverage assessment on neural network based digital twins for autonomous control system, Annals of Nuclear Energy 182 (2023) 109568. URL: https://www.sciencedirect.com/science/article/pii/S0306454922005989. doi:https://doi.org/10.1016/j.anucene.2022.109568.

[21] B. Hanna, L. Lin, P. Athe, T. Son, N. Dinh, Trusting machine learning in nuclear plant control: A reasoning-

based discrepancy checker, in: 12th Nuclear Plant Instrumentation, Control and Human-Machine Interface Technologies (NPIC&HMIT 2021), 2021, pp. 272–284.

[22] A. Gurgen, N. T. Dinh, Development and assessment of a reactor system prognosis model with physics-guided machine learning, Nuclear Engineering and Design 398 (2022) 111976. URL: https://www.sciencedirect.com/science/article/pii/S0029549322003272. doi:https://doi.org/10.1016/j.nucengdes.2022.111976.

[23] J. Lee, L. Lin, P. Athe, N. Dinh, Development of the machine learning-based safety significant factor inference model for diagnosis in autonomous control system, Annals of Nuclear Energy 162 (2021) 108443.

[24] L. Lin, H. Bao, N. Dinh, On the formalization of development and assessment process for digital twins, in: Transactions of the American Nuclear Society, volume 123, 2020, pp. 268–271.

[25] L. Lin, H. Bao, N. Dinh, Uncertainty quantification and software risk analysis for digital twins in the nearly autonomous management and control systems: A review, Annals of Nuclear Energy 160 (2021) 108362.

[26] U. N. R. Commission., Regulatory Guide 1.203 : transient and accident analysis methods, Book, Online, Washington, D.C., 2005.

[27] W. L. Oberkampf, T. G. Trucano, M. M. Pilch, Predictive Capability Maturity Model for computational modeling and simulation, Technical Report, Sandia National Laboratories, 2007. URL: https://www.osti.gov/biblio/976951. doi:10.2172/976951.

[28] P. Athe, N. Dinh, A framework for assessment of predictive capability maturity and its application in nuclear thermal hydraulics, Nuclear Engineering and Design 354 (2019) 110201. URL: https://www.sciencedirect.com/science/article/pii/S0029549319302109. doi:https://doi.org/10.1016/j.nucengdes.2019.110201, special Issue on TRENDS AND PERSPECTIVES IN NUCLEAR THERMAL-HYDRAULICS.

[29] L. Lin, N. Dinh, Predictive capability maturity quantification using Bayesian network, Journal of Verification, Validation and Uncertainty Quantification 5 (2020) 031001. URL: https://doi.org/10.1115/1.4048465. doi:10.1115/1.4048465.

[30] J. Pearl, Probabilistic reasoning in intelligent systems: networks of plausible inference, Morgan Kaufmann Publishers Inc., 1988.

[31] R. Paul, I.-L. Yen, F. Bastani, J. Dong, W.-T. Tsai, K. Kavi, A. Ghafoor, J. Srivastava, An ontology-based integrated assessment framework for high-assurance systems, in: 2008 IEEE International Conference on Semantic Computing, IEEE, 2008, pp. 386–393.

[32] A. Avizienis, J.-C. Laprie, B. Randell, C. Landwehr, Basic concepts and taxonomy of dependable and secure computing, IEEE Transactions on Dependable and Secure Computing 1 (2004) 11–33. doi:10.1109/TDSC.2004.2.

[33] A. Adadi, M. Berrada, Peeking inside the black-box: a survey on explainable artificial intelligence (XAI), IEEE

access 6 (2018) 52138–52160.

[34] E. Tjoa, C. Guan, A survey on explainable artificial intelligence (XAI): Toward medical XAI, IEEE transactions on neural networks and learning systems 32 (2021) 4793–4813.

[35] A. Gurgen, Development and Assessment of Physics-guided Machine Learning Framework for Prognosis System, Ph.D. thesis, North Carolina State University, 2021.

[36] J. Lee, Development of the Machine Learning-based Safety Significant Factor Inference Model for Advanced Diagnosis System, Ph.D. thesis, North Carolina State University, 2020.

[37] B. Nuseibeh, S. Easterbrook, Requirements engineering: a roadmap, in: Proceedings of the Conference on the Future of Software Engineering, 2000, pp. 35–46.

[38] J. Lane, J. Link, J. King, T. George, S. Claybrook, Benchmark of GOTHIC to EBR-II SHRT-17 and SHRT-45R Tests", Nuclear Technology 206 (2020) 1019–1035. doi:10.1080/00295450.2019.1698896.

[39] L. Wang, Development and Application of Data Coverage Assessment for NAMAC Trustworthiness, Ph.D. thesis, North Carolina State University, 2022.

[40] BayesFusion, LLC, Genie modeler, 2021. URL: http://www.bayesfusion.com/.

[41] I. M. Sobol, Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates, Mathematics and computers in simulation 55 (2001) 271–280.

[42] V. G. Weirs, J. R. Kamm, L. P. Swiler, S. Tarantola, M. Ratto, B. M. Adams, W. J. Rider, M. S. Eldred, Sensitivity analysis techniques applied to a system of hyperbolic conservation laws, Reliability Engineering & System Safety 107 (2012) 157–170. URL: https://www.sciencedirect.com/science/article/pii/S0951832011002717. doi:https://doi.org/10.1016/j.ress.2011.12.008, SAMO 2010.

[43] BayesFusion, LLC, Smile engine, 2021. URL: http://www.bayesfusion.com/.

[44] The Assurance Case Working Group, Goal structuring notation community standard (version 3), 2021.

[45] D. Nešić, M. Nyberg, B. Gallina, A probabilistic model of belief in safety cases, Safety science 138 (2021) 105187.

[46] L. Lin, A. Gurgen, N. Dinh, Development and assessment of prognosis digital twin in a namac system, Annals of Nuclear Energy 179 (2022) 109439. URL: https://www.sciencedirect.com/science/article/pii/S0306454922004698. doi:https://doi.org/10.1016/j.anucene.2022.109439.