Trustworthiness of Digital-Twin-based Automation Technology in Nuclear Power Plant Operation (Extended Abstract)

Paridhi Athe[1], Nicholas Cittadino[1], Trang Tran[2], Jacob Yoder[2], Nam Dinh[1], Tran Cao Son[2]

[1] North Carolina State University, Raleigh, NC 27695-7909
[2] New Mexico State University, Las Cruces, NM 88003
pathe|nrcittad|ntdinh@ncsu.edu ttran|jyoder|stran@nmsu.edu

Abstract

This extended abstract describes a collaborative project that seeks to establish a technical basis for, and preliminary development of, a trustworthiness assessment framework for automation enabled by digital twin (DT) technology.

Advanced Nearly Autonomous Monitoring and Control System (NAMAC)

Digital Twin (DT) technology (see, e.g., (Moi, Cibicik, and Rølvåg 2020; Haag and Anderl 2018; Boschert and Rosen 2016; Grieves 2014, 2006; Glaessgen and Stargel 2012)), together with advanced machine learning techniques (particularly, deep learning with physics-informed neural networks), are instrumental in the development of an *advanced **n**early **a**utonomous **m**onitoring **a**nd **c**ontrol system* (NAMAC) aims at enabling nuclear plant's safe, optimal, flexible, and resilient operations. As demonstrated in (Lin et al. 2021, 2022), NAMAC could be considered a viable technology for the development and assessment of flexible plant operations and generation in current light-water reactor nuclear power plants (NPP) in integrated energy systems.

In essence, NAMAC is a computerized safety case that aims to achieve an alignment of Nuclear Power Plant safety design, analysis, operator training, and emergency management by *furnishing recommendations to operators for effective actions* that will achieve particular goals, based on the NAMAC's knowledge of the current plant state, prediction of the future state transients, and reflecting the uncertainties that complicate the determination of mitigating strategies. Figure 1 depicts the transition from operator-centric plant control architecture to NAMAC-enabled plant control architecture.

Trusthworthiness of NAMAC and Beyond

As with any AI-enabled system, NAMAC is (and should be) scrutinized by the question of "*why should an operator trust its recommendations?*". This calls for the development of a framework for the assessment of the trustworthiness of NAMAC, or more generally, for systems that employ DT technology. The trustworthiness of NAMAC is an important

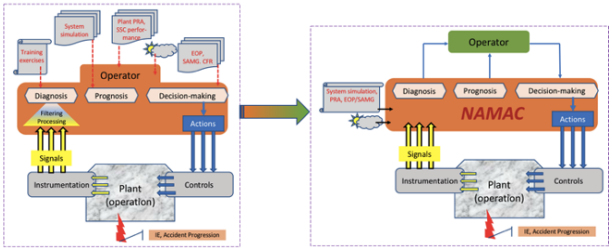

Figure 1: Transition to NAMAC-enabled plant control architecture

question of concern. The issue concerning the trust in NAMAC is partially addressed by the discrepancy checker in NAMAC. The discrepancy checker acts as an uncertainty manager and alerts the operator under situations that are outside NAMAC's scope (beyond the training domain of DTs). However, the previous implementations of discrepancy checker (Lin et al. 2021; Hanna et al. 2021) were purely rule-based with limited knowledgebase (selective rules regarding DT constraints and bounds) and explainability. To address this issue, we focus on developing an improved implementation of the discrepancy checker that focuses on providing operators with explanations on *why does the system recommend action a in the context c?* This is motivated by the fact that explanations can provide transparency of the system, and thus, create trust between the operator and the system. We investigate methods for generating explanations of NAMAC's recommendations using large language model (LLM). This is because

- operators (of an NPP) are knowledgeable in their domain, and therefore, they can accept a recommendation or interact with the system easier if they understand how the system reaches a certain decision;
- LLM can generate natural language sentences and precise natural language explanation helps improve the communication between operators and the system.

In our recent work (Athe, Lin, and Dinh 2024), we tried to enhance the descripancy checker function by adding an LLM component to it. Retrieval augmented generation (RAG) and finetuning were used to enhance the knowledge of LLM in the context of NAMAC. LLMs are stochastic and

their response is governed by a variety of factors (their inherent parametric memory, contextual information in the user query, retrieval mechanism and embedding methods used in RAG, quality and quantity of data used in the fine-tuning, etc).

Taking into consideration the weaknesses of LLM (e.g., weak reasoning ability and hallucination), in this work, we experiment with a pipeline that generates explanations for NAMAC via LLM using a knowledge graph. The pipeline is similar to TranspNet proposed by Machot, Horsch, and Ullah (2024) which integrates reasoning about ontologies with RAG and LLM. This pipeline consists of (*a*) generating a knowledge graph from the knowledge base used in the construction of the DT-modules of NAMAC; (*b*) developing a module that utilizes the knowledge graph and the RAG framework (Lewis et al. 2020) for query answering (QA). The knowledge graph enhances the contextual depth and helps increase the accuracy and relevance of the LLM-generated explanations for NAMAC recommendation. We believe that this will allow for the development of a QA system that takes into consideration the complexity relationships between entities in the nuclear engineering domain and provides explanations with different level of granularity. Furthermore, the lesson learned though this project can be useful for future development of domain-specific QA-systems in high stake applications.

References

Athe, P.; Lin, L.; and Dinh, N. 2024. Using Generative AI to implement the Discrepancy Checker for a Nearly Autonomous Management and Control System for Advanced Reactors. In *2024 International Congress on Advances in Nuclear Power Plants*. Las Vegas, NV.

Boschert, S.; and Rosen, R. 2016. Digital Twin—The Simulation Aspect. In *Mechatronic Futures: Challenges and Solutions for Mechatronic Systems and their Designers*, 59–74. Cham: Springer International Publishing.

Glaessgen, E.; and Stargel, D. 2012. The digital twin paradigm for future NASA and US air force vehicles. In *53rd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference 20th AIAA/ASME/AHS Adaptive Structures Conference 14th AIAA*.

Grieves, M. 2006. *Product lifecycle management*. Nova Iorque, McGraw-Hill.

Grieves, M. 2014. White Paper, vol. 1.

Haag, S.; and Anderl, R. 2018. Digital twin – Proof of concept. *Manufacturing Letters. Industry 4.0 and Smart Manufacturing*, 15: 64–66.

Hanna, B. N.; Lin, L.; Athe, P.; Tran, S.; and Dinh, N. 2021. Trusting Machine Learning in Nuclear Plant Control: A Reasoning-Based Discrepancy Checker. In *The 12th Nuclear Plant Instrumentation, Control and Human Machine Interface Technologies (NPIC&HMIT 2021)*. Virtual, Online.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; tau Yih, W.; Rocktäschel, T.; Riedel, S.; and Kiela, D. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, volume 33, 9459–9474.

Lin, L.; Athe, P.; Rouxelin, P.; Avramova, M.; Gupta, A.; Youngblood, R.; Lane, J.; and Dinh, N. 2021. Development and assessment of a nearly autonomous management and control system for advanced reactors. *Ann. Nucl. Energy*, 150: 107861.

Lin, L.; Athe, P.; Rouxelin, P.; Avramova, M.; Gupta, A.; Youngblood, R.; Lane, J.; and Dinh, N. 2022. Digital-twin-based improvements to diagnosis, prognosis, strategy assessment, and discrepancy checking in a nearly autonomous management and control system. *Ann. Nucl. Energy*, 166: 108715.

Machot, F. A.; Horsch, M. T.; and Ullah, H. 2024. Building Trustworthy AI: Transparent AI Systems via Large Language Models, Ontologies, and Logical Reasoning (TranspNet). arXiv:2411.08469.

Moi, T.; Cibicik, A.; and Rølvåg, T. 2020. Digital twin based condition monitoring of a knuckle boom crane: An experimental study. *Engineering Failure Analysis*, 112:104517.