IAEA NUCLEAR ENERGY SERIES No. NP-T-X.XX

DEPLOYMENT OF ARTIFICIAL INTELLIGENCE APPLICATIONS FOR

V1.11 (TM version)

12 February 2024

INTERNATIONAL ATOMIC ENERGY AGENCY

VIENNA, 2023

FOREWORD

The IAEA's statutory role is to "seek to accelerate and enlarge the contribution of atomic energy to peace, health and prosperity throughout the world". Among other functions, the IAEA is authorized to "foster the exchange of scientific and technical information on peaceful uses of atomic energy". One way this is achieved is through a range of technical publications including the IAEA Nuclear Energy Series.

The IAEA Nuclear Energy Series comprises publications designed to further the use of nuclear technologies in support of sustainable development, to advance nuclear science and technology, catalyse innovation and build capacity to support the existing and expanded use of nuclear power and nuclear science applications. The publications include information covering all policy, technological and management aspects of the definition and implementation of activities involving the peaceful use of nuclear technology. While the guidance provided in IAEA Nuclear Energy Series publications does not constitute Member States' consensus, it has undergone internal peer review and been made available to Member States for comment prior to publication.

The IAEA safety standards establish fundamental principles, requirements and recommendations to ensure nuclear safety and serve as a global reference for protecting people and the environment from harmful effects of ionizing radiation.

When IAEA Nuclear Energy Series publications address safety, it is ensured that the IAEA safety standards are referred to as the current boundary conditions for the application of nuclear technology.

Artificial intelligence (AI) and machine learning (ML) technologies continues to evolve to address different applications within the nuclear industry. There is great interest in harnessing AI/ML capabilities throughout the lifecycle of plant design through the decommissioning process which is expected help transform the industry's economics. This interest is enabled by the technological advancements in areas of sensors, data management, communications, and computational capabilities. As the advancements in AI capabilities continues to uncover applications within the nuclear power industry that were previously undesired or were consider impractical there are several gaps and challenges that needs to be addressed. This would allow the transition the AI/ML technologies from the research and development domain to be used in nuclear facilities. Addressing these gaps and challenges is not straightforward (irrespective of reactor technologies) as there are several considerations and guidance that are expected to influence the deployment of AI/ML technologies. These considerations and guidance would have to address historical safety culture of the industry, several technical aspects, a valid business value proposition, organization and regulatory readiness, and finally end user/public acceptance.

Recognizing the significance and the need to identify different considerations and guidance to support AI/ML technology deployment, the Technical Working Committee comprising of international subject matter experts and advisors from Member States in the frame of the International Network for Innovation to Support Operating Plants (ISOP) was formed. The Technical Working Committee solicited input and reviews from representatives of different Member States and engaged with them during different Technical Meetings to develop this publication. The overall objective of this publication is to capture different considerations and guidance required to support deployment of AI/ML technologies in nuclear power plants. These considerations and guidance are expected to inform Member States in their AI/ML deployment strategy.

The IAEA wishes to acknowledge the valuable assistance provided by the contributors and reviewers listed at the end of the publication, especially the contribution made by V. Agarwal (United States of America). The IAEA officer responsible for this publication was N. Ngoy Kubelwa of the Division of Nuclear Power.

CONTENTS

1. INTRODUCTION

1.1.BACKGROUND

Artificial intelligence (AI) and machine learning (ML) technologies continues to evolve to address different applications within the nuclear industry. There is great interest in harnessing AI/ML capabilities throughout the life cycle of plant design through the decommissioning process which will help transform the industry's economics. Some of the key opportunities for optimal AI/ML application include reactor design, operation and maintenance (O&M) activities, materials, and flexible operations and expanded deployment.

AI/ML has been a topic of research and development within the nuclear power industry for several decades with ebbs and flows of interests. For example, some of the classical applications of AI/ML for reactor O&M includes online condition monitoring using different types of artificial neural networks; fault diagnosis and anomaly detection using unsupervised, semi-supervised, and supervised ML techniques; and maintenance optimization using condition-based monitoring. Technological advancements in the areas of sensors, data management, communications, and computational capabilities (from high performance computation to edge computing) has accelerated the development of new AI capabilities which include (but are not limited to) deep learning, natural language processing (NLP), and more recently large language models. Advancements in AI capabilities continue to uncover applications within the nuclear power industry that were previously undesired or were consider impractical. Some the examples include (but are not limited to) text mining using NLP, integration of predictive analytics with dynamic risk assessment, and automation of several manually performed O&M activities.

Continued advancements in AI/ML technologies and their potential adoption (i.e., deployment) in the nuclear power industry could influence a wide range of needs and applications that could enable the industry to achieve an economically competitive market position within current and future energy markets. What this means for the nuclear power industry is to transition the AI/ML technologies from the research and development domain to their facilities. This transition is not straightforward (irrespective of reactor technologies) as there are several considerations and guidance that are expected to influence the deployment of AI/ML technologies. These considerations and guidance would have to address historical safety culture of the industry, several technical aspects (data accessibility, quantity, quality, governance, technical expertise, explainability and trustworthiness, and cybersecurity), a valid business value proposition, organization and regulatory readiness, and finally end user/public acceptance.

The international Atomic Energy Agency (IAEA) is playing a significant role through The International Network on Innovation to Support Operating Nuclear Power Plants to increase collaboration and experience sharing in the field of innovation for the nuclear industry. AI/ML technologies are one of the innovations within the network scope. For this publication, IAEA is hosting and co-organizing series of meetings where AI/ML subject matter experts, developers, and users representing different member states and organizations (academia, industry, power plants, regulators, and national laboratories) provide their input on different aspects of AI/ML innovations, steps taken to deploy some of the innovations, and lessons learned. This collective knowledge has laid the foundation for capturing considerations and guidance for deployment of AI/ML technologies within the nuclear power industry. These considerations and guidance are dynamic and are expected to evolve with AI/ML technologies. They could also be used to inform a deployment strategy within other areas of nuclear science, such as nuclear fusion and nuclear medicine.

1.2.OBJECTIVE

The objectives of this publication are:

- To provide an overview of various AI applications in the nuclear industry that have the potential to advance the autonomy level (see Section 1.5) when deployed in a nuclear power plant.
- To identify different technical, human factors, cybersecurity, and stakeholders' considerations (including hardware, software, and infrastructure needs) for developing an AI lifecycle system management enabling broader end-user acceptance.

- To emphasize that different categories of data, their relevance, and governance ensuring data integrity plays a vital role in developing an AI technology.
- To consider engaging a regulator and stakeholders at different stages of the development of an AI technology and presenting them with evidence that addresses regulatory, safety, and security requirements, are important for AI technology implementation.

1.3.SCOPE

This publication provides high-level considerations and guidance to enable deployment of AI applications for the nuclear power industry. These considerations and guidance encompass topics related to data, data management and governance; design, development, deployment, operation and quality monitoring of an AI technology; and regulatory, cyber, and stakeholder requirements to implement the AI technology. These considerations and guidance are expected to be adapted as per application needs. It is also noted that AI technology should be used to address specific needs (problem statement) of a nuclear power plant based on the understanding that there could be other technologies or approaches that coul be employed to address those needs. Therefore, it could be best practice to ask the following questions:

- Why is AI needed?
- How will it address the problem of interest?
- What are the capabilities of AI that are better compared to other technological solutions?
- What is needed to develop, deploy, and implement an AI technology?

1.4.STRUCTURE

This publication is organized as follows:

- Chapter 2 provide an overview of wide range of applications and capabilities of AI/ML technologies applied to the nuclear power industry, emphasizing the benefits of the technology.
- Chapter 3 presents insights into different considerations for developing an AI lifecycle system management (design, development, deployment, and operation and monitoring) enabling broader end-user acceptance.
- Chapter 4 discusses different categories of data, their relevance, and governance ensuring data integrity plays a vital role in developing an AI technology.
- Chapter 5 presents different implementation considerations.

1.5.LEVELS OF AUTOMATION

This section provides a brief discussion on levels of automation, as per Section 9 of NUREG-0700 and draws parallel to notional AI and autonomy levels, as per Artificial Intelligence Strategic Plan: Fiscal Years 2023–2027. The United States (U.S.) Nuclear Regulatory Commission (NRC) recognizes the differences between automation and autonomy in nuclear applications that could use AI, as presented in their Artificial Intelligence Strategic Plan: Fiscal Years 2023-2027.

Levels of automation in Section 9 of NUREG-0700 refers to the extent to which a task is automated: Level 1 refers to fully manual (no automation) and Level 5 refers to fully autonomous operation (human monitors performance and perform backup, if necessary, feasible, and permitted). The various levels of automation, as per Section 9 of NUREG-0700, ultimately rely on human-in-the-loop to monitor plant performance and intervene when necessary. Under these guidelines, automation technology[1] aims to provide operator support, rather than replace operator duties in regard to sustained day-to-day operational and tactical control. The U.S. NRC's notional five autonomy levels outlined in its Artificial Intelligence Strategic Plan: Fiscal Years 2023–2027 (see Table 1 of ref), is analogous to the levels of

[1] NUREG-0700 does not refer to any specific automation technology. However, AI can be one of the potential automation technologies.

automation in Section 9 of NUREG-0700, except for two differences identified here:

- Autonomy levels2 specifically calls out the use of AI.
- Level 4: Full autonomous (highest level of autonomy achievable) suggests no human intervention in decision-making compared to NUREG-0700 Level 5 (Autonomous Operation) where human can still intervene.

Both levels of automation and autonomy levels recognize that as technological advancements (including AI) mature and are integrated in nuclear systems for decision-making, day-to-day human involvement is going to reduce. To support different aspects of AI in nuclear discussed in this publication, notional AI and autonomy levels will be the point of reference.

1.6.VALUE PROPOSITION

AI is one of the technologies that is addressing different aspects of nuclear power plant operation and maintenance activities along with other aspects of nuclear fuel cycle. AI has presented immediate value proposition that is strengthening nuclear economics in the competing energy market and increasing operational efficiency by automating some of the manually performed tasks, reducing human errors, enhancing structures, systems, and component reliabilities, enabling predictive maintenance, outage optimization, preventive maintenance optimization, and even nuclear safety. Despite the impressive value proposition of AI/ML technologies in different nuclear applications, it’s adoption in the nuclear industry is slow and face several challenges. This might be due to lack of clear considerations and guidance. This publication attempts to capture some (not all) considerations and guidance, as AI is an evolving as your read this publication.

2. BENEFITS FOR NUCLEAR POWER PLANTS

Artificial Intelligence (AI) & Machine learning (ML) are not new to the nuclear industry. Indeed, there has been considerations on AI for decades, where we can find several interesting publications about AI in Nuclear Power Plants (NPP) published in the 1980. Here is an extract of the introduction of one of those [4] that states the following: “…There was a strong impression that application of this technology (AI) to nuclear power plants is inevitable. The benefits to improved operation, design and safety are simply too significant to be ignored. This is a much different conclusion than might have been reached a few years ago when the technology was new and people were struggling to understand its significance…. It has moved from being a topic understood only by specialists to a situation where users are the most active people in the field.”

Currently, AI/ML covers a broad range of possible applications across an NPP. These include robotics and maintenance, alarm and signal validation, emergency response, process diagnostics, human-machine interface, plant control systems, equipment diagnostics, operation analysis, plant operations and support, probabilistic risk assessment, teaching & learning, fuel manufacturing, etc. These applications provide many possible benefits to the NPP, with cost reduction and safety improvement being some of the most relevant ones. Recent advances in AI research have made today’s AI/ML technologies more mature, accessible, and cost-effective. At the same time, it is fair to say that AI is still an emerging technology to NPPs and most proposed applications are still in pilot phase.

AI provides opportunity to automate tasks to reduce human effort and enables applications that were not achieved before. Often, this is because the task parameters are inexplicable to the extent required for traditional programming. With ML, the automation can be “trained” to find the optimal solution, given sufficient data. Even when the task is or can be automated using traditional means, the data driven models can often be used to improve performance.

Safety and economics are the key issues for the survival and development of nuclear power, which are comprehensively determined by the design, construction, operation and retirement stages of the entire life cycle of nuclear power plants (NPPs). Among them, operation is an important stage which is directly related to the economic benefits. Operation and maintenance (O&M) accounts for significant

[2] Autonomy can be achieved without AI but this publication is not going to cover it.

portion of operating costs. Therefore, there is enormous potential for improving NPP economics by reducing O&M costs. AI technologies are expected to enhance the decision-making and the control capabilities of NPPs through deep mining and utilization of O&M data and the use of innovative solutions.

In terms of safety, by applying the new generation of industrial intelligence technologies, early abnormal situations can be detected, human error can possibly be reduced and complex information can be fused to assist with, and possibly make decisions. In terms of economics, the application of industrial intelligence technology can improve the availability of plants, reduce comprehensive maintenance costs and reduce dependence on personnel, improving the overall efficiency of O&M.

Based on efficient and secure access to the operation and management data of NPPs, the application of AI technology can bring benefits to the safety, the availability and the economy of nuclear power plants, such as:

- Increasing automated operation and auxiliary decision-making, reducing the workload of power plant operators;
- Strengthening the monitoring and analysis capability of equipment status, system status and unit status, improving the safety of power plants;
- Providing more efficient and flexible tools for power plant operation and management personnel.

In summary, to advance the utilization of AI technology with enduring and practical advantages for NPPs, it is important to delineate the scope of AI application with a goal of enhancing plant safety, reliability, and economics. In the following chapter, some current or upcoming applications are reviewed to provide concrete examples of how AI/ML can be used and how it can provide value within the nuclear sector.

.

2.1.OVERVIEW OF GENERIC AREAS OF CAPABILITIES

The range of different AI/ML techniques is vast. For many tasks, there may be multiple different ML approaches that give viable solution. As is the case in many fields of engineering, we find that for any given task a preferred approach emerges, ussually representing the current state of the art.

In the following, the main areas and capabilities are reviewed to help navigate the various ML solutions. Often, the chosen approach significantly impacts the process throughout, from required training data and computing capacity to the available results.

Shallow models (decision trees, linear regression, support vector machines, Bayesian models, etc.) have seen examples, in process optimization, for example. The reported advantages include improved process efficiency and reaction to anomalies as well as early detection of faulty conditions. However,machine learning methods based on Bayesian inference allow explicit uncertainty assessment. Nuclear utilities are looking to quantify real time risk as a function of evolving equipment degradation and ,in the presence of uncertainty, it is helpful to formulate the diagnostic problem in a probabilistic setting in which the quantitative diagnoses are in the form of probability distributions. Adopting a continuous posterior probability distribution allows for more informed monitoring and fault detection decisions compared to threshold monitoring, which tends to have higher false alarm rates. From the posterior distribution, one can compute the probability that a fault has exceeded a certain severity threshold for which a warning or a maintenance intervention action can be made. The Bayesian inference calculations of the fault posterior distributions can be performed from the Markov chain Monte Carlo sampling method. This approach has been shown to be effective for instances of equipment degradation in the feedwater system using archived utility data [10].

Computer vision and deep learning models (convolutional neural networks, etc.) have shown promise in various machine vision tasks from inspection data evaluation, to fire safety improvement and as part of robotic tasks. The reported benefits include improvements in consistency and efficiency of repetitive data evaluation tasks, which due to their nature may tend to be susceptible to human errors. Existing data for cameras and sensors have been used for additional monitoring capability using automated condition detection.

Convolutional neural networks are being introduced in the light water reactor industry to generate more accurate 3-D visualizations of core performance by integrating high fidelity simulations with online sensor measurements. The combined spatial map exhibits reduced uncertainty in important core performance parameters such as linear heat rate and burnup allowing more energy to be obtained from a core loading and with that, increased revenue.

Deep learning networks have successfully been used in evaluation of data from non-destructive testing processes. In service inspections generate vast amounts of data, most of which does not contain indications of flaws. ML models can sieve through this data and pinpoint areas of interest for further evaluation by the human inspector allowing for focus on anomalous data, reducing potential human-factors related errors. Furthermore, the results are available quickly, which helps scheduling outage activities.

Natural language processing (NLP) and the new large language models (LLMs) present a new opportunity with capability to process and produce textual data. Recently, these models have become widely available and have shown potential in nuclear use cases. Amongst other things, the models have shown the capacity to process large textual data sets and provide efficient ways to extract information from or check compliance with textual guidelines. LLMs can also provide context specific guidance and help support best practices, especially with changing workforce.

NLP is being used to explain in human understandable terms how an automated reasoning algorithm arrived at its answer for the case where domain knowledge is embedded and operated on in the reasoning process. In thes instances, the user would like to validate how the algorithm got its answer, but in a form more accessible than the mathematical abstraction inside the reasoning engine. The key enabling insight is that the human has an analogous albeit qualitative understanding of the information embedded in the digital twin. In this application NLP operates on the internal mathematical reasoning sequence and transforms it into a textual description that is the qualitative analogue. This approach is attractive as it brings the algorithm to the human rather than the other way around.

AI/ML models can be as simple as a linear regression model to predict thermal losses of a plant based on cooling water temperature, or as complex as a deep learning model to extract specific insights from a text document. What these two extreme examples have in common are a set of historical training datasets to teach the model the past behaviour of the system and the request for the model to predict the outcome of the system based on other known parameters.

Following the AI/ML community journey, a vast number of applications in past decades were focused on structured data, which loosely translates to numerical data or any type of data that follows a structure, e.g. a table format. Examples include the use of anomaly detection for equipment reliability through clustering and nearest neighbour techniques, the use of classification techniques such as decision tree and support vector machine to categorize health and safety events and the use of non-linear prediction through Neural Network to predict power plant accident conditions and to assess potential future work.

The application of AI/ML to unstructured data, e.g., text data, pictures, videos and audio files, started with the use of traditional NLP to extract keywords from text documents. With recent advancements in NLP and the creation of LLMs, there is a huge opportunity for utilities to harness insights from information within their text documents and databases. As well, advancements in robotics and image processing techniques have created an opportunity for the industry to leverage these techniques for anomaly detection, visual inspection and process optimization.

2.2.APPLICATIONS

Artificial intelligence/machine learning (AI/ML) can automate data processing and interpretation, rapidly analysing vast amounts of data produced by plant operations. Using machine learning algorithms, the AI/ML can identify patterns, trends and anomalies more quickly and accurately, leading to more timely and informed decision-making.

In the following, some distinct example applications are described. The number of currently applied cases is low but rapidly increasing. It is noted that many of the cases shown here are still in pilot phase.

2.2.1. Artificial intelligence in nuclear to enhance learning

Large language models (LLMs) are used in nuclear teaching and learning activities for both internal employees and external clients/stakeholders. This is particularly relevant to the nuclear industry due to specific qualification requirements and challenges with knowledge retention due to transitioning work force.

The use of AI and in particular generative AI could significantly enhance the learning experience and results in the nuclear industry. AI can also improve the cost-efficiency of training activities, by reducing the time needed to produce content and to assimilate it. AI can be used to provide a more personalized, effective and engaging learning experience.

Connecting LLMs capabilities to existing knowledge base may allow people to find answers to their questions quickly and easily. This will reduce the time and effort required to search for information. Here are some of the main use cases tested so far:

- Summarizing training documentation and provide “learning pills” and key take-aways.
- Producing ad-hoc exams using different kinds of questions (multichoice, true/false, fill the blank, etc.).
- Producing new perspectives by relating concepts together or presenting the information in another way.
- Providing specific information for the user’s need, e.g. locating the right insight within large documentation.
- Adapting its responses to the users’ needs, simplifying for instance complex topics for beginner or providing more expert descriptions for an advanced learner.
- Providing multi-language interactions, interaction for instance in Spanish or French on some English documentation and vice versa.

2.2.2. Monitoring and diagnostics: using artificial intelligence to enhance equipment reliability

With recent advancements in data analytics capabilities, the combination of on-line sensor data and continuous models operating on that data have become more available. This is sometimes referred to as a “digital twin” to emphasize the mirroring of the physical assets with virtual models. By integrating real-time data with a simulated environment, digital twins enable predictive maintenance, real-time monitoring, and simulation of future states. In addition to sensor updates and physical or data-driven models, historical information, such as operators logs, engineering inspection and condition assessment reports, any related previous root cause analysis, can be used to update the health of the current asset.

M&D analysts work closely with Operations and Engineering to assess the incoming alerts from anomaly detection models. Like other ML applications, the output of the models need to be verified. These groups are also involved when new ML models are being built to provide their inputs on what instrumentation signals need to be included in the model and what is their respective normal operating values for model training.

The financial benefits that utilities gain through implementing AI/ML M&D cover multiple categories:

- *Cost reduction:* detecting anomalies and acting on them before they cause outages or derates.
- *Time saving:* facilitating access to data and trend analysis to save time for troubleshooting and reporting purposes, optimizing prevention maintenance and condition based maintenance.
- *Avoid generation loss:* first principal models and data driven models can be used to identify areas of potential generation loss and how to improve them.
- *Safety benefits:* Continuous feedback afforded by automated evaluation can improve response and improve safety.

2.2.3. Control rooms and plant control system

As already established in some research and power reactor simulators [50] and in the petroleum industry [51], an advanced alarm system based on expert systems greatly assists the operator by avoiding an overwhelming number of alarms. It works with alarm groups, guides the operator toward a solution, and allows them to navigate to the root of the problem. It is important, following international regulations [52], that the operator has a way to navigate through grouped alarms to access the complete list of triggered alarms.

The control rooms of modern nuclear plants typically have a similar physical layout. In front, there is monitoring of key control variables for both the primary and secondary systems. On the left side, there are physical protection systems, and on the right side, there is a closed-circuit TV system. Operators perform their tasks at workstations consisting of a chair and a minimum of four screens, with two screens used for navigating between different plant systems, one displaying the alarm system, and the last one showing the plant's help manual.

Regarding control rooms, the primary role of AI is to guide the operator in safely operating the plant, assisting them. Therefore, it is possible to replace the help manual (the fourth screen) with an AI assistant that uses gestures and natural language processing techniques.

2.2.4. Text analytics using large language models

Utilities possess a large volume of information in text format, which includes but is not limited to station condition records (SCR), operational experience (OPEX), work order details, maintenance reports, design and operating manuals for various components and systems, periodic inspection reports, root cause analysis reports, etc. There is also a large volume of research and development (R&D) reports that are generated through external entities and agencies. The knowledge that is captured through these documents is useful for day-to-day operation and maintenance, health and safety, equipment reliability, asset management, etc. Most of these data sources exist in siloed databases with limited traditional Keyword search capabilities. As a result, information retrieval from these sources is very labour-intensive, causing these rich resources to be heavily underutilized.

Advancements in NLP through Deep Learning algorithms have resulted in creation of LLM. Capabilities of LLMs include but are not limited to: extraction of insights from the documents, summarization of documents, semantic search through documents, generating reports using multiple documents, etc. It should be noted that expansion of cloud computing provides the hardware resources required by LLMs.

Leveraging these LLMs has become one of the priorities in various industries, including electric utilities. LLMs are available through either open-source codes or APIs through various technology providers. Leading utilities are utilizing these technologies to enhance their internal search capabilities for document retrieval and trending previous events, to gain insights from their previous health and safety related observations and OPEX and to summarize large volumes of text information to a size that is manageable to be read by individuals.

To make these LLMs usable across an enterprise, user friendly web application interfaces are needed to take the input from the user, run the models in the background and provide the desired outcome to the user. There needs to be a close interaction among the business line, the data scientists behind LLMs and the software engineering team creating the user interfaces.

2.2.5. Assistance to non-destructive evaluation inspections

Non-destructive evaluation (NDE) inspections are an integral and important part of nuclear power plant operations; multiple components are periodically inspected for fitness-for-service against known and expected active degradation mechanisms. While some inspections are performed while the plant is in operation, many are part of the busy schedule of a refuelling outage. Inspections can be performed under challenging environmental conditions, while some can generate large volumes of data for posterior review and analysis. Often, data analysis can involve long reviews that require a great deal of attention to and focus on monotonous data, where fatigue and distractions are significant human factors affecting quality and reliability. Other inspections are performed only at long intervals and maintaining high proficiency is challenging.

While the context of NDE inspections vary considerably, at large they are a prime example of an area that can greatly benefit from the assistance of AI tools. There are multiple ways in which AI may assist NDE inspections; the two main ones are:

— AI tools may aid the inspector by providing augmented data for real-time evaluation during live inspections. Such assistance is expected to minimize re-looks or re-scans (where the inspector needs to re-collect data), which can lead to considerable time savings in data collection and, for outage inspections that affect critical path, be a significant advantage.

- AI tools may assist the inspector in post-inspection review and analysis of the data. In itself, this may take different forms:
  - The AI tool may provide fully automated solutions.
  - The AI tool may provide assisted analysis, where it flags regions of interest or potential indications in a larger volume of data that require review by an expert analyst, thus effectively screening the data while maintaining the final decision as a responsibility of the expert analyst.

In either case above, the AI tool may provide assistance for detection, characterization or both. There are additional benefits for the AI tool, having accomplished one or both of those tasks (the “hard” tasks), may provide; these include automation of clerical tasks such as reporting or verification of inspection parameters. These tasks often consume considerable time of the expert staff, thus employing these tools could provide considerable benefit. Furthermore, in the instances of NDE inspections traditionally performed by inspection vendors, such tools can be valuable in enabling efficient and meaningful site staff oversight of the results that would otherwise often be limited because of the high demands on the utility’s NDE staff during outages.

Typical or potential benefits that can be obtained from leveraging AI to assist in NDE inspections include:

- Results are produced faster, allowing the utilities more time to address potential issues.
- Reducing the number of highly qualified personnel that need to be available for the inspection.
- Enabling efficient oversight of vendor activities and results.
- Automation of clerical activities.
- Increased reliability in inspections.

NDE activities that can greatly benefit from AI tools include multiple ultrasonic, visual or electromagnetic examinations performed on multiple components on nuclear power plants. These will typically employ deep networks.

Specific examples currently under development that have reached the stage of field trials include ultrasonic examinations of reactor vessel upper head penetrations (applicable to pressurized water reactor vessels) and dissimilar metal welds ([58]-[60], [63], [64]).

2.2.6. Fault diagnosis and predictive maintenance

AI can be used for continuous monitoring of the operating status of NPP systems and equipment, enabling the rapid detection of potential issues and equipment health management. Such applications reduce plant downtime caused by system and equipment failures, hence enhancing plant safety and economy.

Existing fault diagnosis process rely heavily on expert knowledge, resulting in low efficiency and accuracy. Hybrid data-driven and knowledge-driven solutions enable prompt detection of equipment failure and automatic identification of causes. Such a diagnosis system learns from operational data and fault cases over time, gradually improving its diagnostic accuracy and forming a more intelligent fault diagnosis system, to detect and respond to abnormal situations of NPPs earlier.

At the same time, through continuous online monitoring (OLM), health assessment and lifetime prediction of equipment, preventive maintenance can gradually evolve to predictive maintenance enabling improvement in equipment reliability and maintenance cost which some utilities are already realizing through the transition of time-based to condition based maintenance for various plant equipment and instruments [55].

2.2.7. Operational and decision support

AI based operational decision support, leverages real-time data and prediction models to optimize operational strategies for higher efficiency. With extensive operational data and previous experience, the system can quickly detect abnormal operating conditions, collect and send all information related to these abnormalities, and tune the parameters of controllers. This functionality facilitates effective emergency response suggestions for operational personnel. Ultimately, it assists operators in making prompt and accurate decisions, thereby reducing the risk of incidents.

Vilim et al. [10] describe an AI-based operator aid technology to assist nuclear plant operators to maintain situation awareness and detect faults earlier than would be possible using conventional control room technologies. In practice the process of an operator diagnosing a fault and then adhering to the corresponding paper-based procedure to recover from the fault is time-consuming and prone to error. The objective of this effort was to improve plant performance by addressing plant and grid upset events that presently result in transients that can challenge the protection system. There are safety and economic advantages to reducing the probability that an incident will lead to an unplanned shutdown. The system was implemented on a full scope simulator and a human factors assessment that involved two licensed reactor crews was performed to evaluate crew performance. While the operator response was positive, the technology has not advanced further given the present regulatory framework.

2.2.8. Human performance optimization tools for operators

Operators play a critical role in the operation of NPPs and their actions directly impact the safety of the plants. AI technology can be used to support operator actions to reduce human errors.

Data-driven time series prediction, combined the operator's current operation intention and plant historical operation data, can provide the operator with auxiliary information on operational risk. AI and extended reality (XR) can assist operators in confirming operation steps, receiving the status of equipment, identifying and guiding the operation intention, and providing objective operational hints. With the support of AI, operators performance and accuracy can be improved, ensuring safer operation of NPPs.

2.2.9. Sensor state online monitoring

AI can be used for the online monitoring (OLM) of sensor state, i.e., online estimation of sensor accuracy or response time and diagnosis of sensor health condition during plant operation. OLM can increase the availability of NPP instrumentation and control systems in that if one or few sensors degrades or fails, OLM could be able to provide measurement estimation of the plant process. OLM enables condition-based sensor maintenance management and avoids unnecessary calibration, thus reduces maintenance time and the workload and accumulated dose of maintenance personnel.

2.2.10. Nuclear safeguards applications

AI could be applied in various applications related to nuclear safeguards. AI models such as Artificial Neural Networks (ANNs) have been used to determine the 239Pu content in Spent Nuclear Fuel (SNF) based on simulated signatures of the Self-Indication neutron resonance densitometry (SINRD) [15]. In addition, decision trees, k-nearest neighbours (kNNs) and ANNs were also tested in the identification of the percentage of replaced fuel pins in SNF assemblies based on simulated data from different non-destructive assay (NDA) techniques, namely, the partial defect tester (PDET), the fork detector (FDET) and the SINRD [22]. And in a further step, ANNs were also tested on their capabilities of identifying the presence/absence of individual fuel pins inside SNF assemblies based on simulated measurements of neutron flux and gamma emission rates [17]. The encouraging results indicate the potential for AI in the field of safeguards inspection and the detection of possible diversions in SNF assemblies.

Linear, non-linear and deep learning AI models have been used to predict SNF characteristic parameters such as burn-up (BU), initial enrichment (IE) and cooling time (CT) based on simulated signatures such as gamma-ray intensities and total Cherenkov light intensities [18]–[22]. The identification of SNF parameters is a central task for safeguards inspectors in order to verify the completeness and correctness of operator declarations. Traditionally, this task relies on analysing data from one NDA instrument at a time, however, the utilization of AI models has proven advantageous in their ability to integrate data from different measurement techniques and hence provide more comprehensive results.

AI models have also been considered for safeguards applications in bulk handling facilities. Applying safeguards to such facilities can usually prove to be challenging and resource intensive. Emerging technologies in the field of AI and data science hold promise in enhancing the effectiveness and efficiency of nuclear safeguards in this area. ANNs have been tested for the detection of anomalies in the actinide inventories for a plutonium uranium redox extraction (PUREX) reprocessing facility based on a combination of process monitoring (PM) data and NDA measurements [23]. AI models using such input data can complement the traditional safeguards approaches for bulk handling facilities such as destructive analysis (DA) techniques which are expensive, time consuming and might require on-site analytical laboratories.

2.2.11. Applications for future plants

Future plants may benefit from many of the same usage patterns as existing plants. However, with future plants, the emphasis on design processes and new, sometimes radically different, designs may include new opportunities, in a more native way. At the same time, operational data is not yet available which is a challenge.

The new designs have the opportunity to leverage the power of AI and data analytics as one of the tools to enhance equipment reliability, asset management and life cycle planning. New builds could be instrumented according to the design requirements and the failure mode analyses. The databases could be built in a way in which they are connected to other related databases through data governance principles and the use of AI/ML would follow a holistic view that can benefit the whole organization, rather than isolated practices.

Generating representative operational data for new NPPs is crucial for model training, performance prediction, safety analysis and regulatory compliance. To compensate for the lack of operational data, several techniques can be used. Simulated data from physics-based models can emulate the physical properties and behaviours expected in the reactor. Synthetic data generated through machine learning techniques - al. [67] suggest that by using techniques like generative adversarial networks (GANs), machine learning algorithms can produce synthetic data that mimics the characteristics of genuine operational data. Virtual data based on similar operational settings, such as data from older plants with similar configurations, can be transformed and normalized to act as a stand-in for the new reactor data or used as the basis for machine learning techniques as described above.

AI/ML and deep reinforcement learning has been used to optimize the design process for nuclear fuel. Their AI-based approach could enhance the configuration of fuel rods in a reactor, extending the life of these rods by about 5 percent.

2.2.12. Other applications

A data-driven model for predicting moisture carryover (MCO) in a Boiling Water Reactor (BWR) was constructed using a physics-constrained artificial intelligence technique. An accurate pre-diction of the MCO is of great value for commercial BWR operators as it can be used to modify the operational plan during a power cycle to mitigate high MCO, thereby avoiding elevated dose to on-site personnel and damage to turbine components. This predictive capability is presently in use for the planning of core reloads and for scheduling of operations for cycles already underway. The developers of this technique described the importance of including subject matter expertise on the phenomena being modelled and the importance of understanding the limitations inherent in nuclear data sets, which tend to be “small data”.

AI/ML algorithms are already in use at a number of stations supporting corrective action programs (CAPs) to evaluate the criticality and disposition of reports on conditions adverse to quality. Classification of the severity and disposition responsibility are common and aggregate trending of common features also benefits from AI/ML support. This improves the consistency in annotating the CAP items and reduces the number of hours required for evaluation [61].

AI can help with new material discovery and design, manufacturing and fabrication processes optimization by monitoring and controlling parameters in real time, material behaviour prediction under different conditions and stressors, materials characterization techniques and accelerate it analysis.

Different machine learning and neural network techniques can be used for intrusion detection in the Plant Control System, providing support for diagnosis and predictive maintenance systems. Similarly, these techniques can be used to detect intruders in infrastructure network, and therefore, in each Control Systems of the power plant.

The use of computer vision during events and emergencies to monitor variables at different levels of the reactor building and real-time sensing of radiation data, could be used by expert systems to alert plant operators about potential dangers.

We can also consider that the benefits of AI extend to the world of reactor simulation. For example, it's possible to train a neural network using calculations generated from neutron codes and then use this network as a substitute for the neutronic module in a plant simulator for operator training.

AI/ML can be trained on historical data to provide more accurate outage budget and schedule management. More accurate and realistic schedules provide several potential cost savings including better management of replacement power purchases, improved planning for supplemental outage staffing and better utilization of outage resources. Additional cost savings may result from more accurate identification of supplemental workforce requirements and the timing of additional resources. For example, one utility leverages NLP and machine learning and historical outage schedules to predict schedule logic ties and optimize future outage schedules enabling reduction in human scheduling errors and reallocation of outage work management resources to higher value tasks [56].

AI/ML can process sensor data for radiation levels and provide continuous radiation monitoring and management. It uses machine learning algorithms to learn what normal radiation levels are and can alert human operators immediately when it detects anomalies or danger signals, ensuring immediate response. More accurate dose projections can be made using historical data from previous field work and current plant conditions.

Waste management: AI/ML can be used to optimize the processes for storage, transportation and disposal of low-level nuclear waste. It might be used to coordinate shipments to ensure that transportation and shipping radiation levels are within allowable limits.

Plant Performance Evaluation: Industry oversight organizations are leveraging machine learning algorithms to improve accuracy in predicting plant performance using industry data. For example, the World Association of Nuclear Operators - Atlanta Center have developed models to determine station performance and estimate the assessment score of an NPP using a variety of data such as performance indicators, operating experience reports, assessments, regulatory information and performance trends [57].

2.3.QUANTITATIVE DESCRIPTION OF THE VALUES/ BUSINESS VALUES

The use of AI in utilities provides value and benefits that vary case by case. It can be challenging to quantify the benefits for early AI adoption and, despite clearly identifiable performance gains in specific tasks where AI is applied. The local performance gains may fail to translate to tangible benefits for multiple reasons. It might be that the integration of AI with current processes cause friction, the AI may require full supervision that diminishes the benefits and due to lack of experience with AI, people may discount or fail to capture the various second-order benefits of AI.

Similar to other tools and technologies, the value normally starts small. Once value is successfully proven, the scale-up application would provide business values that are large enough to justify the cost of implementation.

Some noted value from industry during field trials of the technology include ([55], [56], [59], [63]):

- Provideing results (even if preliminary) faster than what is otherwise possible, allowing better time to prepare for or address potential issues.
- Providing valuable help in managing workforce challenges such as availability and maintaining sufficient proficiency for seldomly performed activities.
- Reducing the number of highly qualified individuals required to be on site. Besides leading to decreased costs, this also helps with workforce availability.

Other potential benefits include:

- *Increased operational efficiency,* such as allowing efficient oversight and faster issue identification.
- *Increased operational reliability*: AI tools can help with known human factors issues, such as those associated with fatigue and distraction.
- *Enhanced health and safety (H&S):* AI can be used to enhance H&S for the employees and contractors by learning from previous events and providing insights from historical incident reports. The insight can be extracted through numerical AI/ML or through LLM techniques. These

algorithms can highlight the important trends and identify hidden trends that are hard for humans to capture, due to the large number of data points. As with other AI tools, human oversight is needed to review the AI models outputs before implementing findings. Other information, such as Observations, if they are collected, can be used to trend as a precursor to H&S incidents.

- *Enhanced financial and operational performance*: financial and operational performance trend together. AI/ML could help utilities move from a labour-centric preventive maintenance strategy to a technology-driven predictive maintenance strategy. M&D Centre and digital twin (DT) applications, explained above, are examples of how utilities can achieve better performance. AI/ML can also be leveraged to minimize repetitive activities organizations do in their daily job. Enhancements in supply chain, IT services, financial and performance reporting are a few examples in which AI and automation techniques would save a lot of time for employees to be spent on more productive activities. Application of Robotics, which comes with integrated AI, when implemented correctly would enhance operational and financial performance. Performing routine visual inspections, performing work in high radiation dose areas to minimize staff exposure time (this ties back to the H&S benefits above as well) and deploying the robots to capture picture/video/audio data as needed, are just a few examples of how application of AI can enhance operational and financial performance.

2.4.LESSONS LEARNED

While the experience with deployed AI/ML solutions is still limited, some common success factors can be identified and lessons learned, which characterize many of the application examples.

There’s currently quite high interest in AI/ML and the new opportunities offered by this technology. At the same time, the wide general media coverage around these technologies may build expectations or concerns that are not pertinent to the specific application being considered. Experience has shown that it is important to facilitate sufficient communication among the stakeholders to explicate the expected benefits and limitations of the AI/ML application and to focus on possible concerns relevant to the application.

The role of subject matter expertise and domain knowledge in conjunction with the AI/ML expertise is emphasized as a key success factor on several applications. Experience with engineering systems is that AI methods are made more robust and reliable if domain knowledge is embedded in the algorithm, in this case physics information. Many engineering problems in the light water reactor industry are high value but are resistant to solution by traditional engineering methods, e.g., [3]. ML coupled with subject matter expertise can provide a solution to these otherwise intractable problems. In the case of AI/ML NDE data evaluation, the role of subject matter experts, i.e. inspectors, have proven decisive in developing successful models with high reliability. The role is important both in preparation of high quality training data during model development and then in evaluation of the model performance in the application context.

The development of the NDE data analysis has shown the importance of building the models and applications incrementally and providing the users opportunity to build trust in the models. Throughout the development, EPRI has reported the results in industry events and shown applicability in field trials. Utility personnel, inspection vendors and regulator representatives have had the opportunity to witness and monitor the development and raise potential issues to be resolved prior to adoption. As the models near field application, the role of industry guidance increases. Especially with a new technology, like AI/ML, the value of best practices and industry level guidance, such as EPRI MRP documents [65] and European network for inspection and qualification (ENIQ) recommended practices [66] help individual licensees adopt the new technologies in a safe and controlled way.

3. AI SYSTEM LIFECYCLE MANAGEMENT

3.1.INTRODUCTION

Current and anticipated uses of AI in the nuclear industry extend beyond targeted data analysis of defined scope datasets. The latter cases have a short shelf life and likely rely on a small number of stakeholders. An approach is hence needed to address the following aspects:

- The solution development may need to integrate input from interdisciplinary teams of technical specialists, AI developers and conventional software developers.
- The solution may result in a source code base of moderate or large size.
- The outcome of the project may be needed by a large number of stakeholders.
- The developed solution may undergo regulatory scrutiny, as it may produce results of safety or operational significance.
- The developed solution will likely have a long shelf life and require maintenance, version control and documentation.

Lifecycle framing enables envisioning AI applications in a scalable manner. Furthermore, the lifecycle approach ensures that AI applications are designed to meet end-user needs, are developed according to design intent, can be kept operable during their intended lifetime and consider regulatory needs during the initiation phase.

The remainder of this chapter aims to identify topics that may warrant careful consideration as part of the AI lifecycle and explain their significance. The chapter also provides a high-level overview of some approaches and tools that address the identified topics, to serve as an illustration. The following is not intended as a standard specification or AI development methodology but a set of guidance principles based on the prevalent state of knowledge within the nuclear industry at the time of writing.

3.2.DESIGN

3.2.1. Defined AI problem statement and requirements

As part of the problem statement, it is often necessary to convert a problem or task into a set of AI-based requirements. In some cases, this is a simple process that replicates a human decision (e.g., deciding on whether a crack exists in an image). However, in many scenarios, problem statement formulation can involve a complex process of interpretation of domain-specific requirements and translation into a machine learning problem statement. For example, ongoing efforts are researching means to introduce automation into compliance verification in plant inspection activities. Inspection procedures are extensively prescriptive and provide detailed requirements. Converting such requirements into a set of well-defined data-driven decisions and the associated metrics requires careful review and analysis of a subject matter expert involvement that has experience in both the inspection procedure and AI.

Applying AI approaches requires reframing the problem into desired outcomes of the AI solution, based on the need for:

- Classification of labelled data (e.g. fault classification).
- Prediction of a response variable as a function of predictor variables (e.g. remaining useful life of a component as a function of monitored parameters).
- Insights on any inherent clusters or groupings in the data.

3.2.2. Technical basis

A technical basis for the use of a particular AI methodology to address a given problem needs to exist prior to implementation within an AI application. The nature of a given AI problem statement determines what prior technical basis may exist to allow the use of AI tools.

Some AI problem statements are generic and mature, having a strong technical basis with widely available generic solutions. The development effort cases primarily consist of confirming problem-solution fit and integrating the software elements into a working application. Some examples of AI problem statements with generic solutions are:

- Image recognition: pre-trained deep neural networks exist such as AlexNet are available for these applications.
- NLP: tools such as chatbots can be used to address the majority of applications involving the analysis, processing and generation of textual information.

On the other hand, some AI problem statements are application specific and lack a generic solution. A specific AI methodology needs to be devised to address these problems. The development of the technical basis then becomes part of the development process. Some examples of AI problem statements requiring the development of a technical basis are:

- Developing data-driven surrogate models which can be used instead of computationally expensive codes (e.g. as a substitute for neutronics or thermal-hydraulic models).
- Fault identification/classification and remaining useful life estimation of nuclear plant equipment.
- Fault tolerant approach to substitute faulty sensor readings using artificial neural network in a nuclear plant.

3.2.3. Static and dynamic models

AI models can be categorized as static or dynamic models. Static models are usually models that were created and frozen for direct use (e.g., a classifier that can determine if some text relates to safety event in nuclear power plants) or transferred and augmented to take advantage of the AI model but for a different application. An example of this type of approach (referred to as transfer learning) is the use of transformers, which are models that are trained a large corpus of data (such as Meta's RoBERTa for text) and used in a more limited context. The benefit of using static models is that it reduces the risk and impact of model changes. In the case of transfer learning, it enables the rapid design and use of smaller and computationally fewer challenging models that are faster to train, tune and run. They also benefit from the broader knowledge of the parent transformer. For example, if a classifier is built to identify if some text relates to an electrical event, the word 'wire' could be associated with words such as 'cable' and 'terminal' by the transformer, even though the augmented model might not know those words. The benefit of using static models is that the qualification/validation effort is front-loaded, likely only needed during initial development and unless a new version of the transformer is used or the static model needs to adjust to new conditions, this qualification remains valid.

Dynamic models are more challenging to use in nuclear applications because they change in time and are, therefore, onerous to qualify. This change can be to the structure or architecture of the model or limited to the parameters of parts or the whole model. The change can even extend to transformer models, if one is used. Some layers of those models can be unfrozen to allow tuning for the specific use case. Dynamic models are used when there is an expectancy of process change to the point where the AI model cannot adapt and needs to retrain and retune to track the change. For example, the use of AI in control is challenged by the aging of the controlled process and the model needs to retrain to understand the emergent conditions. From a qualification perspective, dynamic models require development of continuous model validation approaches, or to adopt an incremental retraining and validation approach in time during operation.

3.2.4. Risk reduction

Like any model, AI models are expected to fail. A model that achieves 100% accuracy is practically not feasible and usually indicates a problem with the testing dataset (i.e., either the dataset is too limited and not representative of all system states, or it is biased to mirror the training and validation datasets). Therefore, it is essential to use models with the anticipation that they will fail. In practice, reducing the risk of failure is achieved by two approaches. The first relies on biasing the validation metrics toward the undesired output. For example, if fire detection is used by a camera video feed, a false positive (fire is flagged when there is no fire) is much more tolerated than a false negative (a fire occurs and is not detected). Therefore, the F-Score is usually of higher order than the typical F1 score to bias the metric used toward the missed detection. The second approach relies on biasing the loss function or threshold used in the training and classification process, respectively. It is possible to add a penalty for a specific false classification or prediction to convince the model that it is more important to detect a certain class than another. However, this improvement of one class or prediction often results in the degradation of others.

3.2.5. Human factors considerations

AI/ML applications are likely to involve human users at some point in their lifecycle. Therefore, human factors need to be considered in the design. For instance, if the AI/ML is used for an operator support system or an automatic system, the interaction between operators and the AI/ML-based system should be carefully designed.

Systems may fail when the design does not consider the interaction between operators and system. Many issues related to human performances in the interaction, especially with the automation, have been reported, including the followings [82]:

- Out-of-the-loop unfamiliarity;
- Clumsy automation;
- Automation-induced errors;
- Inappropriate trust;
- Inadequate training and skill loss.

The out-of-the-loop unfamiliarity means reduced ability of the operator to detect automation failure and to resume manual control [83]. This issue often happens in a highly automated (or autonomous) system in which the operator is removed from directly performing the control. "Clumsy" automation refers to the automation making easy tasks easier and hard tasks harder. This occurs when easy tasks are automated and hard tasks are still left to the operator. A typical example is mode errors in which the operator takes an action when the automation is in the wrong mode. Inappropriate trust includes misuse and disuse. The misuse means the operator over-relies on the automation and so fails to detect the failure of automation and to intervene during the operation, while disuse refers to the situation in which the operator does not use the automation because of the low trust in its reliability. Inadequate training and skill loss refers to situations in which the automation may eliminate the opportunities for the operator to obtain skills by doing the job. More issues and explanations are presented in [82].

One approach to address the issues above is human AI (HCAI). HCAI attempts to build on user observation, operator engagement, usability testing, iterative modification and continuous evaluation of human performance [84]. It aims to augment or enhance human performance. To be human-centred, an AI/ML should address fairness, accountability, interpretability and transparency [85]. To do this, Xu suggested a framework of HCAI that consists of three components, i.e., ethically aligned design, technical enhancement and human factors design [86]. The ethically aligned design means that AI/ML solutions should avoid discrimination, maintain fairness and justice, and not replace humans. Technical enhancement addresses that AI/ML technology should be enhanced to reflect the depth characterized by human intelligence. Lastly, human factors design should be adopted to ensure that AI/ML solutions are explainable, comprehensible, useful and usable to human operators.

Cooperative AI, i.e., another concept to design the interaction, imposes more emphasis on the collaboration between human and AI/ML, while HCAI is more concerned with improving human performance. AI/ML research has focused on improving the individual intelligence of agents and algorithms. On the other hand, cooperative AI focuses on improving social intelligence which is the ability of groups to effectively cooperate to solve the problems they face [87]. It covers the broad range of cooperation that includes AI-AI, human-AI and human-human with AI support to facilitate the cooperation. Dafoe et al. suggested four key capabilities to for cooperation such as [88]:

- *Understanding:* the ability to take into account the consequences of actions and predict another's behaviour, belief and preference;
- *Communication:* the ability to share information explicitly and credibly with others in order to understand behaviour, indentions and preferences;
- *Commitment*: the ability to make credible promises if needed for cooperation, and;
- *Norms and institutions:* social infrastructure, e.g., shared beliefs or rules, that reinforces understanding, communication and commitment.

3.2.5.1.User acceptance

Introducing AI into the nuclear industry requires careful consideration to address concerns and build trust among practitioners and users. Strategies AI practitioners may employ to make users more comfortable with AI in the nuclear industry include:

- Transparency and explainability: Provide clear explanations of how AI algorithms work and their decision-making processes. Transparency aids users to understand the rationale behind AI recommendations. Using interpretable models and avoiding "black box" approaches may make it easier for users to trust AI outcomes.
- Robust validation and testing: Conducting rigorous validation and testing processes to enhance reliability and accuracy of AI models can help demonstrate the effectiveness of AI systems and address concerns about their performance.
- Collaborative development: Involving experts and stakeholders throughout the development process. Such collaboration helps the AI solutions align to end-user needs. making practitioners more comfortable with the technology.
- Education and training: Providing comprehensive training programs to educate users about AI technology, its capabilities, and its limitations empowers practitioners to work effectively with AI and reduce anxiety associated with the unknown.
- Risk assessment and mitigation: Conducting thorough risk assessments to identify potential challenges and vulnerabilities associated with AI implementation and developing strategies to mitigate risks and ensure the safe and secure use of AI technologies in nuclear applications may help increase user acceptance.
- Incremental implementation: Introducing AI technologies gradually and in controlled environments allows practitioners to observe and understand the impact of AI on specific tasks, making the adoption process more manageable and potentially less intimidating for the user.
- User involvement in decision-making: Involve end-users throughout the decision-making process of AI technologies, soliciting feedback, addressing concerns, and incorporating user input may enhance the acceptance of AI.
- Ethical Considerations: Emphasizing the ethical use of AI and clearly communicating the ethical principles guiding AI development and deployment may help to reassure users that AI technologies are adhering to ethical standards.
- Continuous monitoring and improvement: Implementing robust monitoring mechanisms to track the performance of AI systems over time. Regularly updating and improving AI models based on feedback and evolving demonstrates a commitment to ongoing improvement and enhancement to the user.

Addressing these considerations may assist AI practitioners in fostering a culture of trust and acceptance, facilitating the successful integration of AI technologies.

3.2.6. Designing for the cloud

There is significant interest in leveraging cloud infrastructure for AI applications. The main difference when designing AI applications for the cloud lies in the infrastructure and in its ability to scale: the memory and computing power allocated in a cloud-based software solution increases or decreases with traffic, which is useful for managing variable workloads and guaranteeing high availability.

Currently, cloud software is generally based on containers, which are isolated, portable environments that can be managed by orchestration platforms such as Kubernetes. Containers are used by software developers to run their code locally in an environment that contains the same content (operating systems, libraries, settings) as the one where their code will be executed in the application. Several software elements may have different requirements and are often run within different containers in a single application.

These technologies are useful to data scientists in the nuclear industry, enabling better reproducibility. The cloud infrastructure relies more and more on code and less on manual procedures to limit human mistakes: we talk of infrastructure as code. The infrastructure design itself is part of the design of the cloud-based AI application but could be adjusted more easily during the development phase if necessary.

3.2.7. Robustness and resilience

While the previous sections have mostly focused on developmental challenges, AI models face significant deployment challenges in the face of adversarial threats that actively seek to undermine their performance. Unlike past decades where physical models were closely guarded secrets, with open-source learning and the wide availability of effective AI algorithms, adversaries can mimic physical models of critical infrastructure by simply observing and relaying the data passing through the system to train their own models [30]. Furthermore, they can detect and exploit weaknesses in the model by exploiting various correlations among the data to send false signals to systems and misguide them. Specifically, with the GAN framework, it is possible to generate synthetic data closely approximating the underlying distribution of true data and the problem is considerably simplified if the adversary possesses domain knowledge, as is the case with state-sponsored and advanced persistent threats. For example, an adversary may alter the steam flow rate in a nuclear reactor but falsify the sensor measurements with synthetic data to bypass detectors and make it appear as if the system is under normal operation.

Furthermore, adversarial examples have also been generated by harnessing the blind spot of DNNs via seemingly imperceptible perturbations [38]. As mentioned earlier, DNNs do not necessarily learn relevant features and may be activated through minor perturbations of irrelevant features that exploit the gradient-based learning algorithms in networks. These perturbations, while appearing noisy to the human eye, cause DNNs to completely misclassify samples and make erroneous decisions. While initial research perturbed the entire input data, recent research has demonstrated that it is possible to change only a handful of input features such as a single pixel on an image to cause a serious misclassification.

Significant work has been done in recent years to improve the robustness of AI to adversarial examples, e.g., robust encoding, sparse representations, pre-processing of input data to remove adversarial inputs [40], [41] etc. However, parallel advancements on the adversarial side include membership inference attacks [42] that may be constructed to decipher the training data used to create the model. Poorly trained models that overfit the training data are especially prone to such attacks.

A key ingredient missing in the above discussion on robustness and resilience is that of “common sense” [48]. Common sense is a difficult concept to quantify, let alone express mathematically and may be loosely defined as the aspects of intelligence that most humans take for granted or conclude subconsciously without much thought. Achieving robust and resilient intelligence requires AI agents to exhibit common sense through learning causation, intuitive physics and the ability to reason. Currently, AI is very far from this goal and even large systems with hundreds of billions of parameters [49] do not appear to show signs of artificial general intelligence or common sense.

3.2.8. Cybersecurity principles

While there are numerous potential applications of AI for nuclear power, there are computer security concerns that must be addressed. Conversely, AI has the potential to improve nuclear power facility computer security. An overview of these matters is presented, as well as highlights of how IAEA guidance can be applied to ensure the computer security of computer-based systems, which incorporate AI, for nuclear facilities.

Computer security Objective: First, it is important to determine the computer security objective. IAEA NSS 17-T Rev. 1 [73] defines computer security – a synonym for cyber security – as “a particular aspect of information security that is concerned with the protection of computer-based systems.” For computer security, protection relates to ensuring the confidentiality, integrity, and availability of computer-based systems and information – the CIA triad. According to IAEA NSS 23-G [74], these properties (or objectives) can be defined, as follows:

- Confidentiality: The property that information is not made available or disclosed to unauthorized individuals, entities or processes.
- Integrity: The property of accuracy and completeness of information.
- Availability: The property of being accessible and usable upon demand by an authorized entity.

Overall, the objective of computer security for nuclear facilities is to ensure that cyber-enabled adversaries do not compromise nuclear safety, nuclear security, and nuclear material accountancy and control (NMAC) functions through an attack on a supporting computer-based system.

- *AI Computer security vulnerabilities:* The use of AI can result in new vulnerabilities in computer-based systems at nuclear facilities, including those performing functions related to nuclear safety, nuclear security, and NMAC. If a malicious actor would seek to exploit such vulnerabilities in order to compromise the functions performed this could lead or contribute to a nuclear security event. An incomplete summary of AI-specific threats and vulnerabilities is presented below. A knowledge base of known adversary tactics and techniques to AI systems can be found in the MITRE ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems) framework [75] when possible, references to the framework are provided.
  - *Confidentiality:* In some cases, artifacts that are associated with AI could contain sensitive information (Section 1.1 [74]) – these artifacts include AI models and training data, for example. An adversary may attempt to steal this information in various ways (AML.T0035). This can be achieved via "conventional" cyber-attacks that seek to exfiltrate data (AML.T0025) and attacks wherein the adversary performs targeted queries of an AI model in order to gain insights about the – potentially sensitive – data that was used to train it (AML.T0024).
  - *Integrity:* Ensuring the integrity of AI is of the utmost importance – critical decisions that are related to nuclear security and safety could be informed by predictions made by AI models. There are several attacks that are specific to AI that target integrity. Attacks have been demonstrated that allow so-called backdoors to be placed into AI models that are introduced during model training; these backdoors elicit prescribed output from a model when nefarious input is provided (AML.T0018). This is a form of model poisoning (AML.T0018.000). A related attack are so-*called adversarial examples, wherein an adversary provides inputs to a model that are manipulated – imperceptibly* to the human eye for images, for example – to cause a model to make a targeted or untargeted misprediction (AML.T0043). These attacks have been demonstrated, for example, for AI that can be applied to security surveillance systems [76] and computer antivirus software [77].
  - *Availability:* In a similar manner to other systems that do not use AI, those that make use of the technology are susceptible to attacks that can compromise their availability (AML.T0029). This often takes the form of so-called denial of service (DoS) attacks, wherein voluminous service requests are made causing a target to become overloaded and unavailable for legitimate use. In the case of systems that incorporate AI, this type of attack could manifest as large volumes of requests to classify input data or inputs that are intentionally crafted to cause excessive resource use, causing the systems that host the classifier to exhaust computational resources. As many AI applications require specialized computing hardware, such a graphics processing unit (GPUs), it may not be possible to readily scale resources to mitigate these kinds of attack.
- *Adversarial use of AI:* In the same way that nuclear facilities are considering the use of AI to make their operations more effective and efficient, adversaries are doing the same. This nefarious use of AI can take many forms and may include the use of AI to support the rapid development of malware capabilities, learning the behaviour of a target facility to enable more effective engagement, the generation of deep fakes to subvert surveillance systems and generate uncertainty, and the automation of malicious activities. Concretely, we are seeing evidence of the use of LLMs to generate more realistic phishing emails; a relatively straightforward use of this technology, but one that could have significant impact (AML.T0052). The result could be an adversary that is more effective, which can operate at larger scales with lower costs. The consequence for nuclear facilities and the sector is a potentially significant shift in the nature of the risk that needs to be understood and addressed.
- *Defensive use of AI:* Conversely, the security community has invested in uses of AI to improve computer security. For example, machine learning models have been used extensively to support the detection and classification of malicious behaviour that manifests in network or host data. These models are trained on datasets that either represent normal – for anomaly detection – or malicious behaviour – for malicious behaviour classification. A significant challenge the sector faces is a shortage of skilled computer security professionals; to help alleviate this problem, AI is being applied to automate security tasks.

- *Securing AI Systems:* Despite the new computer security challenges that AI introduces, there is a significant body of research and guidance that can be used to manage the risk – these can be broadly categorized into approaches to improve the security and robustness of AI itself and techniques that seek to ensure the computer security of AI in the context of a system. For example, there are techniques that can be applied to training models that are used for AI in an adversarial setting that have the purpose of making them robust to adversarial attacks [78]. Federated learning is an approach to enabling privacy-preserving machine learning, which could be applied in this context to reduce the risk of disclosing sensitive information (confidentiality) [79].

Many of the recommendations for computer security best practice and guidelines can be used to protect systems that use AI. For example, [73] advocates the use of a graded approach to computer security, wherein "[…] the strength of computer security measures put in place to protect a facility function is in direct proportion to the potential worst case consequences of a compromise of the facility function." (Section 2.26 [73]). Use of this approach implies that systems, which use AI, that support higher criticality functions should attract stronger security requirements; this could entail limiting the use of AI, due to the computer security risk and the consequences associated with compromise, and applying stronger security controls to systems, as part of a facility's Defensive Computer Security Architecture (DCSA) (Section 4.67 – 4.82 [73]). This could, for example, involve implementing strong computer security measures [80] to systems that execute AI models, which support more critical facility functions, with the objective of preserving the correct performance of those functions. It is important to note these measures should be technical, organizational, and physical in nature. The security of the supply chain is a concern in relation to AI – attacks, such as model poisoning, can be realized via the supply chain. The IAEA has published guidance on securing the supply chain [81] that can be applied to manage this risk, including placing security assurance requirements on suppliers while ensuring that such procurements do not introduce unresolved risks to nuclear security.

To summarize, there are computer security considerations that need to be understood in relation to the use of AI at nuclear power facilities; these can include new vulnerabilities and adversarial uses of the technology that change the nature of nuclear security risk assessments. There are additionally uses of AI to support computer security. There are significant volumes of best practice, standards, and guidance that can be applied from the IAEA and other organizations. These are important and should be applied using a graded approach. There are open computer security questions about the use of AI for nuclear power facilities, including answering fundamental research questions about the vulnerabilities they introduce and how they can be mitigated, along with questions regarding how computer security assessments should be made for regulatory approval of systems that use these technologies.

3.3.DEVELOPMENT

3.3.1. Software quality assurance

Software in nuclear safety, safety-related, or safety-significant applications is typically developed according to defined software quality assurance (SQA) standards, as part of overarching quality assurance (QA) programs. AI applications with safety significance are likely to fall within the scope of such SQA standards.

The goal of QA standards and programs is to set up systematic and controlled processes for software aspects likely to affect its performance. Such areas are typically the procurement, development and use of software and digital systems.

3.3.2. AI solution qualification

Software for nuclear applications requires qualification to verify that it is fit for use. Qualification is an integral part of the development lifecycle and is intended to assess the technical basis, verify the implementation and validate the performance of software against "real-world" data. Software qualification requires a systematic and documented approach, which is typically defined a priori.

If an AI application is anticipated to have operational or safety significance, software qualification may be a good process for reducing the likelihood of failure. Pursuing such an effort is expected to provide clarity to contributors regarding performance expectations and would lead to enhanced end-user satisfaction.

3.3.3. Repeatability

AI models are reproducible if, under the same training setup, the same result is reached under the same evaluation criteria. Some machine learning approaches, such as deep learning, incorporate different sources of randomness such as initialization of random values in neural network weights, random selection of which data are used for training and which for testing, hyperparameters tuning. Introducing different levels of randomness during training is essential for building robust and performing models but can be a problem when trying to reproduce the results achieved. While some levels of randomness in the training stage, such as the selection of which data to use for training and which to use for testing, can be handled simply by setting different seeds, parallel processing and GPU-specific libraries can include different levels of nondeterminism that can affect reproducibility.

3.3.4. Independent verification, validation and uncertainty quantification

A major challenge associated with the use of open-source data and models is they may compromise independence. Often, testing data used for validation is from open sources and the models used are also using open-source tools and libraries. Because open-source datasets overlap and models introduce systematic errors, it is challenging to claim independence when performing validation of AI models. Two approaches can be adopted to overcome this. The first relates to ensuring independence by limiting the testing datasets to a pre-screened set of data that is qualified for benchmarking. However, if those benchmarks are published, model developers could use them as a validation set, compromising the validation independence and biasing the model performance to address the benchmark dataset. Another approach is to accept the lack of independence and design the validation process to assume a certain level of dependence. Tools can be created to determine the level of dependence of datasets.

3.3.5. Open source vs. proprietary software

AI solution development involves several choices related to the use of proprietary or open-source tools. Such choices may relate to:

- Programming languages.
- Software libraries and toolboxes.
- Training data.
- The AI algorithm.

For each item above, there exist open-source and proprietary alternatives. At a high-level, the advantages and disadvantages can be summarized as:

- The use of open-source libraries is low cost and enables access to a vast amount of pre-developed software. However, it can create version control challenges and lead to uncertainty around verifiability and quality.
- The use of proprietary tools provides QA traceability, which is preferred (and in many cases a requirement) in nuclear applications. However, proprietary software often involves high costs and a small developer community.

3.3.6. Internal/external development

Nuclear organizations may choose to develop software in-house or engage external partners to conduct software development activities on their behalf. Software proponents may have no prior AI software development experience, or no software development experience. In such case, the software proponent may involve a partner to provide the required expertise. In such cases, they should ensure that:

- There is a clearly stated AI problem statement, or that a preliminary scoping phase is planned to obtain one.
- The availability and characteristics of the training data are known a priori (covered in Chapter 5 in further detail), such as:
  - Quantity and frequency of availability.
  - Format and mode of access.
  - Labelling.
  - Adequate coverage of the variance in the possible sample space to be usable.
  - Category (e.g. synthetic, mock-up, laboratory, field).
- Performance requirements are clearly stated.
- Any applicable SQA requirements can be met, whether working under the proponent's QA program or as a member of the supply chain.

3.3.7. Development process

Agile development is an iterative and flexible approach that emphasizes collaboration, customer feedback and incremental progress. Regular product releases are encouraged to deliver value to users sooner and get their feedback on all new features. Adjustments are possible throughout the project and the priorities can change. In contrast, waterfall software development is a more linear approach where the project is made up of phases, each starting when the previous one is completed. Customer input is mainly gathered at the beginning of the project and the product is delivered all at once after a lengthy development phase, with possible discrepancies between what is delivered in the end and the true needs of the users (that may differ from the specifications that may be written at the beginning of a project).

When developing an AI solution, the agile development process gives access to the end-user's reactions to the AI results, making it easier to know whether further tuning is required or if additional features are needed to better understand the results. However, some time-consuming algorithm developments may require longer development cycles before being presented, to avoid disappointment for the user.

Key AI features of interest to the nuclear industry include but are not limited to the following:

- *Continuous learning:* AI systems can learn from data and experiences, improving their performance over time. Machine learning and deep learning are common approaches where algorithms adapt and evolve based on input data.
- *Reasoning:* AI can use logic and reasoning to make decisions. This involves processing information and drawing conclusions based on established rules and patterns.
- *Problem solving:* AI systems excel at solving complex problems by analysing data, identifying patterns and making informed decisions. AI can process and analyse data at a much faster rate than humans. This is crucial in situations where real-time decision-making is required to maintain safe and efficient plant operations.
- *NLP and speech recognition:* AI can understand, interpret and generate human language. This enables interactions between humans and machines through natural language, enabling applications like virtual assistants.
- *Perception:* AI systems can perceive and interpret their environment through various sensors, such as cameras/thermal imaging, acoustics, temperature, humidity and other environmental sensing systems.
- *Adaptability:* AI systems can adapt to changing environments and evolving data. This adaptability is crucial for handling real-world scenarios.
- *Automation:* AI can automate repetitive tasks, allowing for increased efficiency and productivity. This includes robotic process automation and autonomous systems.
- *Machine vision:* AI systems can "see" and interpret visual information, making it possible to analyse images and videos.
- *Learning from experience:* AI systems can improve their performance by learning from past experiences and adjusting their behaviour accordingly.
- *Decision making:* AI can make decisions based on data analysis, optimizing choices to achieve specific goals.

- *Self-learning:* AI systems can continue learning and improving without explicit programming, allowing them to adapt to new challenges.
- *Parallel processing:* AI systems can handle multiple tasks simultaneously, leveraging parallel processing capabilities for faster computations.

3.3.8. Training algorithms and metrics

Training metrics are essential in the development of AI models. These metrics help data scientists and machine learning engineers evaluate and improve their models during the training process. Some key training metrics include:

- *Loss function:* Measures how well the model is performing by quantifying the error between predicted and actual values. Lower values indicate better performance.
- *Training loss and validation loss:* Track the loss on the training data and a separate validation dataset. It helps in identifying overfitting (when validation loss increases while training loss decreases) and aids in model selection.
- *Learning rate curves:* Show how the learning rate impacts the loss function. It helps in choosing an appropriate learning rate for training.

3.3.9. Initial and continuing operator training AI

AI can provide insight and training strategies to address weaknesses in operator performance. Using test and simulator performance of operators, AI can recommend training program improvements, procedures enhancements and MCR human machine interface improvements to improve the performance of operator actions. Data can be collected on crew, individual and department level performance to identify trends and weaknesses in performance. A custom program can be developed to target those weaknesses to improve individual, crew and department performance.

To achieve this AI application utilities will need to build a dynamic learning model of testing data from all initial and continuing written exams, JPMs and simulator scenarios from historical database and continue to use current training data to augment the AI recommendations. The feedback loop will require additional steps in the operations training process of training planning and exam analysis. The dynamic feedback will allow the AI model to continue to be refined with newly identified performance weaknesses and recommendations for targeted training to close those identified gaps.

3.3.10. Accuracy vs. false positives vs. false negatives

These metrics are crucial for evaluating the performance of classification models and are especially relevant in applications like medical diagnosis and fraud detection.

- *Accuracy:* Accuracy measures the overall correctness of predictions, i.e., the ratio of correctly predicted instances to the total instances. While accuracy is a common metric, it may not be suitable when the classes are imbalanced. For instance, in fraud detection where genuine transactions far outnumber fraudulent ones, a high accuracy can be achieved by simply classifying everything as genuine.
- *False positives (type I error):* False positives occur when the model incorrectly predicts a positive outcome when it should have been negative.
- *False negatives (type II error):* False negatives occur when the model incorrectly predicts a negative outcome when it should have been positive.

Balancing accuracy, false positives and false negatives depends on the specific use case and the associated costs or consequences of each type of error. Sometimes, optimizing for one may come at the expense of another.

3.3.11. Accuracy, recall, precision, etc.

- *Accuracy:* As mentioned earlier, accuracy is the ratio of correctly predicted instances to the total instances. It provides a general measure of model performance but can be misleading when dealing with imbalanced datasets.
- *Recall (sensitivity):* Measures the ability of the model to correctly identify positive instances out of all actual positive instances. It is particularly important when the cost of missing a positive instance (false negative) is high.
- *Precision:* Measures the ability of the model to correctly identify positive instances out of all instances it predicted as positive. It is essential when there is a high cost associated with false positives, like in spam email detection.
- *F1-Score:* The harmonic mean of precision and recall. It provides a balanced measure of both false positives and false negatives and is useful when balancing the trade-off between precision and recall.

Understanding and choosing the right evaluation metrics is critical in AI model development. The choice of metrics should align with the specific goals, use case and potential consequences of the model's predictions, however it is important to consider the trade-offs between accuracy, false positives and false negatives.

3.4.DEPLOYMENT

3.4.1. Cloud and edge deployment

To Implement a cloud based predictive maintenance system, data capture, data storage and analysis systems and transport infrastructure must be built.

For the on-site facilities:

- A system of preferably, wireless sensors must be in place.
- Transmission protocols must be chosen. ZIGBEE and Bluetooth are two compatible technologies for transport between sensors and the on-site network that are possible. The possibility also exists, as most sensors have this capability, to have sensors directly communicate with the on-site wireless infrastructure using mobile protocols.
- An aggregation network must be created. There will be areas where traffic must be aggregated due to difficulty in transmission through walls and other barriers.
- A transmission backbone must be chosen to aggregate data at the site and transmit it to the cloud. DAS implementations are very expensive and have high maintenance overhead. Mobile Private Network (MPN) costs and a less expensive, more flexible option. MPN's are inherently cheaper and more flexible than DAS networks with the advent of a new spectrum not controlled by the traditional carriers. Specifically, Citizens Broadband Radio Service (CBRS) refers to the unlicensed spectrum in the United States that can be leveraged for private 5G or 4G LTE networks. It consists of 150 MHz of spectrum in the 3.5 GHz band. In this case, not only are first costs lower but as there is no need for wireless connectivity to the wireless carrier and thereby very expensive mobile charges are avoided.

Cloud side requirements:

- Data management software will need to be selected to organize and provide easy access to the wide variety of data in the system and be convenient. Historical as well as current data will be stored.
- Model retraining software will need to be hosted as well as a version control code repository to host the analytical software.
- User Access will be by web interface. The user, from their desktop, will have the ability to find data, hosted on the cloud and load that data into various software modules for analysis and reports.
- User Validation will be provided by the system referencing utility employee records.

Backups, variable capacity and site maintenance are all provided by the cloud host.

The recent developments in edge computing have made it a viable alternative to cloud based deployment. Edge computers are small, energy efficient computers optimized for data processing and intended to be deployed near the data source. Despite their small size, edge units can include significant computing power and are often optimized for AI/ML computing loads. Thus, they can be used even in applications with quite high computational requirements.

The primary benefit of edge deployment is that processing data near the source reduces the need for data transfer and the distributed units promote more scalable deployment. In a nuclear context, edge computing offers the additional benefit of providing security by physical separation. The edge units can be deployed as stand-alone appliances, isolated from external networks. This makes deployment of AI/ML in secure environments easier.

Internally, the edge deployment uses largely the same tooling as cloud deployment with the edge unit acting as a small server. Hence deployment can often be easily moved between cloud and edge and either can be chosen based on application needs and requirements.

Relevant examples of edge AI/ML deployments include the recent field trials in NDE data evaluation [72].

3.4.2. Process/change management

The challenges of implementing AI/ML within an organization extend beyond technical and human factors considerations impacting work processes, job functions and organizational culture which are common across industries. Many technological advancements throughout history, such as the invention of the steam engine or the internet, have resulted in automation and digitization in ways that have disrupted workforces [89]. Therefore, in order to ensure investment in AI/ML technology and applications are successful, practitioners should consider pairing their deployments with change management strategies. Process and organizational change management is an important aspect in the general delivery of new technology solutions within the Nuclear Power Industry, with particular considerations and challenges to plan for when implementing AI/ML applications.

While AI/ML driven applications have increased in popularity of late, organizational change management is a well-studied field and it can benefit a practitioner to follow an established Change Management model to ensure they are identifying and addressing the important considerations of change applicable to their area of practice. Some established frameworks include but are not limited to the following:

- Kotter's 8-Step Change Model: a methodical approach for implementing successful organizational change, emphasizing the need for a sense of urgency, strong leadership, a clear vision, and engaging a broad base of stakeholders to drive and sustain change efforts. [90]
-
- Lewin's Change Model: a 3-stage theory of change involving an "Unfreeze" stage to prepare for the change, a "Change" stage where the transition occurs, and a "Refreeze" stage to stabilize and integrate the new state as the norm. [91]
-
- ADKAR Model: this framework includes 5 essential elements Awareness, Desire, Knowledge, Ability and Reinforcement, with a step-by-step approach that emphasizes individual and human-aspects of change. [92]

Regardless of the change management framework or model that one decides to follow, it is important that this should be treated proactively as part of AI/ML solution delivery and integrated into the AI lifecycle to inform design, development, deployment and risk management practices as required. Therefore, assessment of the current state is fundamental for understanding readiness for AI. For example, stakeholder analysis might reveal low proficiency in the technology or a preference for interpretable AI models. In this case developers may elect to design or select algorithms, that may be less complex, but offer better explainability, providing greater clarity and trust to end-users. Some additional common considerations are establishing a clear vision, identifying and analyzing stakeholders, communication, establishing feedback mechanisms and monitoring. Common causes of resistance to change when implementing AI/ML are job replacement and workforce reduction, lack of technology proficiency in staff and managers, lack of trust in AI/ML systems and lack of trust in leadership decision making around AI/ML technologies [93]. However, a change management plan should be tailored to the specifics of both the application and organization and scaled as appropriate based on risk and effort. Finally, the deployment of AI/ML solutions within the nuclear power industry demands a change management strategy that is as dynamic as the technologies themselves. Recognizing the iterative and evolutionary nature of AI/ML technologies, change management is best when integrated into each phase of the AI lifecycle while also remaining agile, and responsive to ongoing feedback and technological advancements. This strategic agility ensures that change initiatives both address the current state and are anticipatory of future developments, underpinning the long-term success of AI/ML applications.

3.4.3. Risks

In 2022, the US Department of Energy released the AI risk management playbook (AIRMP) [5], which is a comprehensive reference guide for AI risk identification and recommended mitigations (actionable pathways) to support responsible and trustworthy (R&T) AI use and development. This guide references more than 140 identified risks. [5] is the distribution of those risks within AI lifecycle, risk type and AI principles (risks can be in several sub-categories):

TABLE 1. DISTRIBUTION OF THOSE RISKS WITHIN AI LIFECYCLE, RISK TYPE AND AI PRINCIPLES [5]

AI Lifecycle	**Risk Type**	**Primary Principle**
Problem Identification ^{52}	Algorithmic ^{70}	Accountable ^{14}
Supply Chain ^{38}	Data ^{63}	Accurate, reliable and effective ^{50}
Data Acquisition ^{62}	Design ^{55}	Lawful and respectful of our Nation's values ^{15}
Model Development ^{77}	Equity ^{25}	Purposeful and performance-driven ^{11}
Model Deployment ^{57}	Ethics ^{37}	Regularly monitored ^{8}
Model Performance ^{54}	Operational ^{65}	Responsible and traceable ^{8}
		Safe, secure and resilient ^{24}
		Transparent ^{5}
		Understandable ^{6}

The following lists some of the inherent risks generally associated with AI applications. It should be noted, however, that the risk of using AI should always be weighed against the associated risk of not using AI. For instance, there is always the risk AI will produce the incorrect answer; but consideration should be given to consider the risk of the alternative, non-AI solution producing the incorrect answer under the same conditions.

- Over- and under- reliance on AI tools: Over-reliance on AI tools may lead to unintended results, including loss of expertise in the field. Under-reliance can render the tool useless and negate any potential benefits it brings.
  Mitigation strategies are often similar for both cases and can include training and designing the combined process in such a way that it maintains the desired level of human agency.
- Unintended or undetected model extrapolation: AI tools can be exposed to input data beyond the characteristics of the data it was trained on. In such cases, the response of the AI model can be erroneous.
  Mitigation strategies include introducing checks on the input data to attempt to detect out-of-scope conditions, monitoring input data for data drifts and monitoring model performance to detect performance degradation.
- AI applications are, in general, built based on historical training datasets: The performance of AI models depends on the quality and quantity of those datasets. The data quality depends on how

much and what information is provided at the time of data entry and if these entries are related to what the AI model is expected to simulate.

- Siloed databases: Another risk facing NPPs are siloed databases, developed over time among various business lines. These datasets, even though they might contain related information, may not be well-connected. Leveraging all the information in these databases becomes more challenging as the number of these siloed databases increase. Close collaboration among various operational components is needed to bridge this gap and minimize the underutilization risk. Implementing data governance practices would help to minimize this risk.
- Deployment risks: There are risks of an implementation and deployment nature and risks that involve installed capability. In the former there is the integration of new capabilities into the existing suite of software in a plant. Generally, utilities are resistant to wholesale change in software but look for incremental change that brings new value with minimal disruption. Once deployed, issues can arise from misapplication of the algorithms for which they were designed, which generally results from any of the following:
  - Lack of understanding of the tool,
  - Failure to maintain currency with the real time process as it undergoes aging, reconfiguration, or other changes,
  - Inability of the algorithm to generalize across multiple process applications, but instead hardwired for each individual application,
  - False positive alarming that leads users to discount the value and from a lack of transparency as to how the algorithm obtained its result.
- Vulnerability assessment and penetration testing: It is important to verify the AI infrastructure using ethical hacking tools such as vulnerability assessment and penetration testing. In the same way, infrastructure monitoring must be carried out to detect anomalies; AI can help with this task by applying advanced detection techniques. The use of AI in nuclear computer-based systems will require the intervention of auditors by Information Security, who can do different types of tests on the software (white box, black box, etc.), and understand how these algorithms are fed, and ensure that their behaviour does not 'degenerate' during the operation.

Perhaps a better method to understand the risks associated with the implementation of an AI application would be to define some industry-wide AI Categories and promote an AI scoring criteria within the nuclear energy industry in order to provide:

- Common understanding of AI worldwide;
- Intuitive reading for non-technical stakeholders;
- Transparent risk evaluation;
- Easier qualification process and regulation for high-level scoring applications.

Another opportunity to reduce risk would be the creation of an open-data ecosystem within the nuclear energy industry. More data will lead to better results and lower risks. In particular, the availability of specific datasets for rare events may enable improvements of AI solutions for NNPs operation.

3.4.4. Guardrails

3.4.4.1. Definition

AI solutions need to integrate with the current operations. How much autonomy the AI assumes and how readily its functioning can be controlled and corrected by human operators differs in different applications. With increasing autonomy, the demands for AI validation and safeguards increase.

3.4.4.2. Parallel and gradual deployment

Given its novelty in the nuclear sector and concerns around AI in general, one practical safeguard that can at times be used is parallel and gradual deployment of AI.

Parallel meaning it is initially and temporarily deployed alongside currently adopted and accepted solutions. In this interval, the existing solution is still the valid or “official” one. Gradual in that different elements of the AI solution are enabled at different times; for example, this may represent a gradual increase in the level of autonomy of the solution.

Such a period of parallel deployment brings several benefits:

- Allows for relevant field testing of the AI solution in a safe environment: any failures would have no consequences.
  - Conversely, differences in the outputs may also bring to light shortcomings in the existing practices that are mitigated by the AI solution.
- Helps build trust in and acceptance of the technology.
- Provides an opportunity under relevant circumstances for the staff to become familiar with and how to properly interact with the AI tool in anticipation of a potential transition.
- Assist identifying any issues in the human-AI interface of the system.

It is important to note that an AI solution must often be intentionally developed in a manner to enable parallel and gradual deployment. This should drive design decisions from data input to model architecture and outputs that may even limit the AI solution. However, the potential end-benefits can justify this decision, especially in that they tend to considerably facilitate adoption and acceptance of the technology. The AI-assistance of ultrasonic non-destructive evaluation inspections is one example where this has indeed been observed.

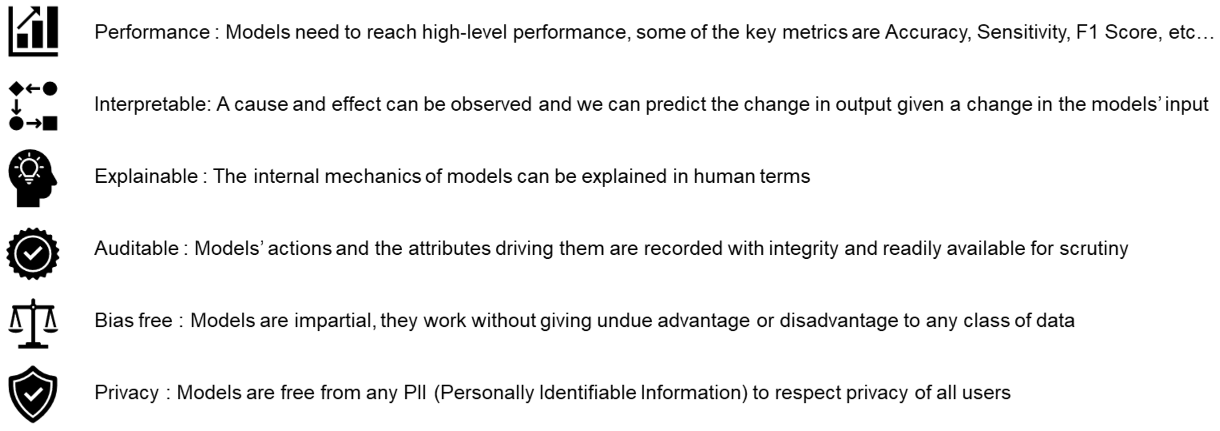
3.4.4.3. Workforce training

Training can be a potential method to minimize issues when deploying AI solutions. This refers more to training users on how to use and interact with the AI tool and how to properly interpret its outputs. It is valuable for users to be informed about anticipated failure modes applicable to the AI tools they interface with and how to recognize those failure modes. It should be noted that AI is not a replacement for domain expertise; users should still have sufficient proficiency in the applicable domain.

Note that training on AI itself may not be or is not necessarily, an effective way to mitigate issues on any particular application. The tool should provide outputs that are relevant to the applicable domain of operation and immediately understandable to staff with expertise in the area and not overly dependent on AI/ML terms or concepts.

“By design” principles

In general terms AI should consider some primary principles “by design”, which means that those principles should be considered at the very beginning of any AI project, in its description or design phase, before starting the development. Those principles “by design” are the following:



Performance : Models need to reach high-level performance, some of the key metrics are Accuracy, Sensitivity, F1 Score, etc…

Interpretable: A cause and effect can be observed and we can predict the change in output given a change in the models’ input

Explainable : The internal mechanics of models can be explained in human terms

Auditable : Models’ actions and the attributes driving them are recorded with integrity and readily available for scrutiny

Bias free : Models are impartial, they work without giving undue advantage or disadvantage to any class of data

Privacy : Models are free from any PII (Personally Identifiable Information) to respect privacy of all users

FIG. 1. Primary principles “by design”

3.4.4.4. Level of autonomy

Even when an AI system is nominally under constant oversight, maintaining human agency and effective oversight may pose challenges in some instances and the models may exhibit greater than intended effective level of autonomy.

The majority of AI applications at NPPs currently reported focus on low autonomy and low level of plant safety significance. Nevertheless, it is of interest to also explore the potential of more autonomous applications.

While the promise of autonomous operation lies in the potential efficiency gains from improved allocation of human resources, the challenge is for the human acting as the ultimate decision authority to maintain a commensurate level of situational awareness. Clearly, as the time needed to respond to events shortens the human supervisory role is increasingly challenged. Hence, the near-term opportunity for AI deployment lies in the part of the continuum where AI provides recommendations for operation and maintenance activities that are addressable on a time scale that allows for human assessment and validation. This takes in automated monitoring and tracking of component conditions and machine recommendations for the highest value action to be taken in case of diagnosed performance degradation. An active area of research involves using a digital twin to sense component degradation and forward this to a Markov decision process [13]. There are pilot projects underway in the U.S. aligned with this objective.

3.4.4.5.Control feedback loop

The next step up the continuum is for the control feedback loop to be closed by the machine but with a provision for the operator to take manual control. One approach to this problem is to preserve the lower-level PID single-input single output controllers in existing NPPs and have their setpoints generated by a supervisory controller that contains AI elements for decision making and for advanced control to enable enhanced performance. By taking manual control the operator remains in the loop and has the ability to issue setpoints and recover the plant.

3.4.4.6.Human oversight

AI solutions and products should be considered as another tool in the toolbox used for decision making. The autonomy level each tool is used at would depend on criticality of the decision as well as the reliability of the AI products. The level of human oversight would also depend on maturity of the AI products and would vary case by case.

To reduce human oversight over time and increase autonomy of AI applications, the 'explainability' of the AI product and application needs to be improved. By being more explainable, AI applications can be more rigorously tested and trusted and therefore would be considered less as black boxes (which need heavier human oversight). This would also assist regulators to review and approve applications in which AI can be used in nuclear applications.

3.5.MAINTENANCE AND QUALITY MONITORING

Successful implementation of AI within organizations operating nuclear facilities requires a team of Subject Matter Experts (SMEs) from affected programs and technology solutions. SMEs will need to work together to develop products that are both feasible and technically correct to minimize adverse impacts on quality and compliance. AI learning requires extensive and accurate input to develop models to meet business needs. SMEs provide/identify this training data and provide validation of model results. In the same spirit, experts in technology solutions must be used to integrate the model into existing solutions or create new ones as needed.

A cross-discipline review of AI performance is required for success. Engaging experts in the technical process will help make sure the desired results are obtained. Early recognition of declines is more likely when knowledgeable individuals are directly involved in the performance reviews of model outcomes. Without these reviews, the viability of an AI system is greatly diminished and it cannot be relied on for critical decisions, thus negatively impacting quality and compliance.

Continued efforts by SMEs must be formalized into the implementation strategy for AI integration. Lack of validation and reinforcement of model outcomes may lead to a decline or stagnation in accuracy as time elapses since implementation. For example, a model trained to classify condition reports in terms of priority and severity may need to be updated as plant staff turns over and differences in writing styles emerge. Initial training data may no longer accurately categorize future condition reports.

Evaluating performance periodically by testing is necessary to ensure results are still as desired. Depending on the level of degradation an entirely new model may need to be trained to prevent correction attempts that lead to overfitting. This degradation should be identifiable through statistical analysis of data, dependent on the application use of the AI. Applications centred around natural language may be harder to assess than those used for analysing data points collected from instruments.

Models that self-train run the risk of overfitting their data through a reinforcement loop. In nuclear applications, this could lead to undesired outcomes and should involve a human component. This may be the role of a QA group or individual departments that are responsible for the process that is being automated via AI.

4. DATA CONSIDERATIONS

Data plays a vital role in AI applications. AI applications depend on data to learn and operate and their success largely relies on the degree to which the related data is accurate, complete, consistent and relevant. Regardless of the quality of an AI/ML algorithm, its results can ultimately be unreliable when using inadequate data.

Typically, data supporting the development of AI/ML applications is divided in three groups based on their use as follows:

- Training data
  Data used for training the AI/ML model. This typically accounts for the larger part of the available data. For supervised learning, this typically consists of curated input/output pairs.
- Validation data
  Data used in combination with the training data with the purposes of tuning untrainable design parameters of the model (called hyperparameters). As an example, the depth of a decision tree is not a trainable parameter in the model; instead, it is set as a design parameter. Validation data enables assessment of the proper choice of such parameters, as well as assessment of the robustness of the model.
- Testing data
  Data used to test the model after it has been trained. This dataset should be completely independent from the training and validation datasets. Any data or information leakage from the training or validation datasets to the test dataset can compromise model performance assessment, biasing it towards more favourable results.

All three data groups above should have similar characteristics (scope, distribution, etc) and should be relevant to the desired application. As will be discussed in Section 4.1, different data sources are more or less suitable for the different uses. The proportion of the total available dataset separated into each of these groups can vary. Typically, the training dataset uses the largest portion of the available data, and the validation and testing datasets are of similar size. ISO/IEC 8183:2023 [2] provides discussion on these data subsets.

FIG. 2 illustrates the use of data through the lifecycle of AI/ML applications. During development, data from different sources are used for model training and testing; during operation, the trained AI models use field data to make predictions or decisions. In applications involving dynamic models, AI models continue to learn and evolve over time by leveraging new field data for further training through a feedback loop (illustrated by the red line in FIG. 2).

It is clear the correctness and reasonableness of AI outputs are highly dependent on the accuracy, relevance and overall quality of the input data. During the life cycle of AI/ML applications, data goes through many processes including generation, transfer, transformation and use. As shown in FIG. 2, data for the development of an AI/ML model for a nuclear power plant (NPP) or other domain application can be collected from various sources (e.g., field, laboratories, simulations, open source). This data can then be directly used or processed, i.e., transformed, to be suitable for the training and validation of the AI/ML model.

Given the importance of data for AI/ML applications and the overall lifecycle shown in FIG. 2, this section discusses the following key topics:

- Data sources:
  Supporting data can (and often will) come from several sources, which can have their own limitations and considerations and can be more adequate for one stage or use than another. The different potential data sources to support AI/ML model development are identified and discussed in the Data Categories subsection.
- Data quality:
  The overall characteristics of the underlying data is crucial for AI/ML applications.
- Data integrity:
  In the life cycle of AI/ML applications, several potentials exist for the supporting data to have deteriorated, be misused, or corrupted, whether accidentally in processing or by intentional attacks.
- Data management:
  The Data Management subsection contains considerations related to the overall management of data through its lifecycle, from data records to quality monitoring and audits and others.
- Data sharing:
  Representative data in sufficient quantity is paramount for the successful application of AI/ML solutions. However, for certain applications of interest in the NPP domain, it is unlikely that any one single plant would have sufficient relevant data to enable a successful solution and therefore data sharing becomes of ultimate importance. The final subsection of this section discusses benefits and challenges of data sharing.

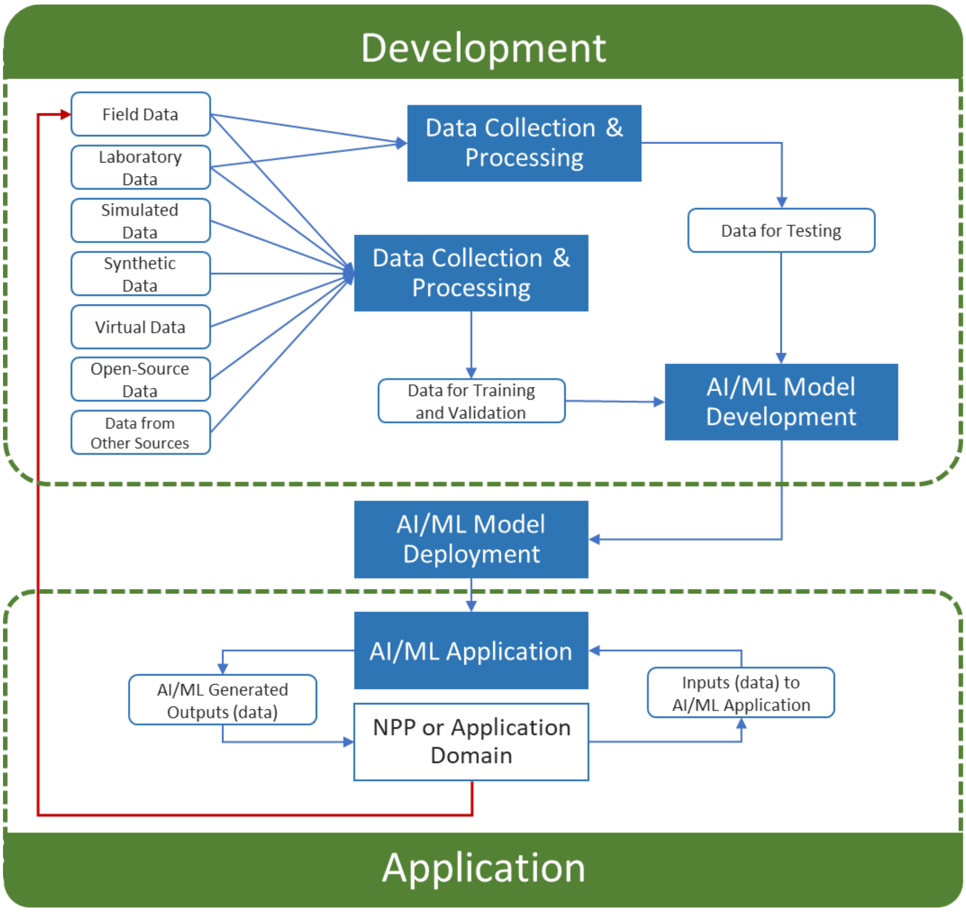


FIG. 2. Data usage through the lifecycle of AI/ML applications

4.1.BENEFITS OF DATA LIFE CYCLE FRAMEWORK

The data life cycle is crucial in AI/ML systems for several reasons:

- Data Quality Assurance: Ensures that the data used for training and inference is accurate, reliable, and representative. A well-managed data life cycle helps identify and rectify issues such as missing values, outliers, and biases, enhancing the quality of machine learning models.
- Model Training Efficiency: Streamlines the process of preparing and feeding data into machine learning models. Effective data collection and pre-processing during the data life cycle contribute to more efficient and faster model training, saving time and resources.
- Model Generalization: The data life cycle aids in creating models that generalize well to unseen data. By incorporating diverse and representative datasets during training, models become more robust and capable of making accurate predictions in real world scenarios.
- Scalability: Efficient handling of large datasets is essential in AI/ML systems. The data life cycle framework supports the scalability of machine learning models by providing mechanisms to manage and process increasing volumes of data.
- Adaptability to Change: Data patterns evolve over time, and models need to adapt to these changes. The data life cycle, when integrated with regular model maintenance, allows for continuous improvement and adaptation, ensuring that models remain relevant and effective.
- Decision-Making Confidence: Reliable data throughout the life cycle provide confidence in the decision-making process. Stakeholders can trust the insights generated by AI/ML systems when they know that the data used is of high quality and has undergone thorough processing.
- Compliance and Governance: Adhering to data life cycle best practices facilitates compliance with regulations and governance standards. It ensures that sensitive information is handled appropriately, mitigating risks associated with data integrity, privacy, and security.
- Resource Optimization: An organized data life cycle minimizes inefficiencies in data management. This optimization is crucial in resource-intensive AI/ML projects where effective utilization of computational resources and manpower is essential for success.
- Continuous Improvement: The data life cycle supports a feedback loop, allowing organizations to learn from the performance of deployed models. Continuous monitoring and assessment lead to iterative improvements, ensuring that AI/ML systems stay relevant and effective over time.
- Data life cycle overview: The data life cycle for AI systems encompasses the processing of data from the earliest conception of a new AI system to the eventual decommissioning of the system and is separated into a number of distinct stages. Each stage will often, but not always, be part of a data life cycle for an AI system.

4.1.1. Data life cycle framework

A data life cycle represents all the stages through which data can pass within any system that uses data of any kind. It is designed to support the achievement of objectives related to system governance, system utility, data quality and data security, by ensuring that data processing is given due consideration during the planning, development, use and decommissioning of the system.

ISO/IEC 8183 defines the stages and identifies associated actions for data processing throughout the AI system life cycle, including acquisition, creation, development, deployment, maintenance and decommissioning.

The data life cycle for AI systems encompasses the processing of data from the earliest conception of a new AI system to the eventual decommissioning of the system and is separated into a number of distinct stages. Each stage will often, but not always, be part of a data life cycle for an AI system.

The data life cycle framework, shown in FIG. 3, identifies a set of conceptually distinct stages that data used in an AI system undergo, from data planning to data decommissioning.

4.2.CATEGORIES OF DATA

As mentioned, data supporting AI/ML can, and often will, come from various different sources. This subsection identifies the main potential sources of data, providing a definition and important considerations for its usage in support of AI/ML applications.

4.2.1. Field data

Field data is collected directly from sensors or human input within the operational environment of a nuclear facility. It represents the only type of data seen by the AI/ML model during operation.

Considerations for use include:

- Can be critical data for real-time monitoring, predictive maintenance and operational decision making.
- Should be validated rigorously to ensure accuracy and safe decision making, especially when using sensor data.
- When used for near real-time inference and decision making, online sensor data should be monitored for calibration drift or periodically checked for calibration.
- Typically require data stewards who are experts in the associated system and component being monitored via instrumentation.
- It can be used at any stage. It is essential that it is the main type of data at final testing. Even if it is not the only type of data in testing, performance should also be assessed exclusively on the field portion of the test data.

4.2.2. Laboratory data

Laboratory data is data gathered in controlled settings to conduct experiments and physical simulations. Examples include inspection data on laboratory specimens.

Considerations for use include:

- Training and validating AI models should be used in conjunction with field data and known physical models for testing.
- It can support initial selection and development of AI models.
- Laboratory data is often clean and well-structured, making it ideal for the initial training phases of AI models; however still requires testing in real world settings.
- Care should be taken to ensure that laboratory data adequately represents the complexities of real-world data.
- Laboratory data can augment datasets that are difficult to collect in real operational settings. For example, in-service failure data is typically rare for critical operating equipment. Lab environments can be used to intentionally fail components to collect data across various failure modes when building reliability and predictive maintenance models. In these situations, the data will still lack operational nuances that only field data can provide.

4.2.3. Simulated data

Simulated data is data generated through computational or physical models to simulate real-world scenarios. Examples include data generated by FEM models.

Considerations for use include:

- Data augmentation can support scenarios with insufficient data quantity or distributions.
- Computational models should be validated against real world data where possible and with rigor commensurate with the risk associated with the use of AI model outputs.

- Simulated data is useful for initial AI model training and validation due to cost effectiveness but generally should not be solely relied upon for testing.
- Special care should be taken to ensure that simulated data adequately represents the complexities of real-world data.
- Not suitable for operational decision making.

4.2.4. Synthetic data

Synthetic data is data generated through methods other than computational or physical models that describe the physical phenomena. Examples include data generated by AI methods (e.g., generative adversarial networks).

Considerations for use include:

- Data augmentation can be useful in scenarios with insufficient data quantity or distributions.
- The data is validated against real world data where possible and with rigor commensurate with the risk associated with the use of AI model outputs.
- Synthetic data can support initial AI model training and validation due to cost effectiveness but should not be used for testing.
- Care should be taken to ensure that synthetic data adequately represents the desired complexities of real-world data.
- Not suitable for operational decision making.

4.2.5. Virtual data

Virtual data is data that is generated by manipulation of existing real data (i.e., field or laboratory data). Examples include manipulation of data to alter relevant characteristics (such as size or location of a target defect).

The key difference between virtual data and synthetic or simulation data is that it is built directly from existing data (field or laboratory data). The motivation is to expand the scope of the data beyond what is available as field or laboratory data while still leveraging real datasets to capture characteristics that perhaps are not as easily reproducible in synthetic or simulated data. For example, in building an anomaly detection model using motor bearing temperature readings, the dataset may lack the full spectrum of operational temperatures. Developers can extrapolate missing temperature data using relationships between existing sensor readings and ambient conditions, creating “virtual data” to cover the full temperature range for comprehensive model training. Further examples can be found in [68]-[72].

Considerations for use include:

- Data augmentation can support scenarios with insufficient data quantity or distributions.
- Virtual data could be used for stress testing AI models under extreme conditions that may be unsafe or impractical to replicate in the real world.
- The data lineage (source and transformations) applied to virtual data should be well-documented and understood.
- Virtual data should be validated against real world data where possible and with rigor commensurate with the risk associated with the use of AI model outputs.
- Virtual data can be used for initial AI model training and validation due to cost effectiveness but generally should not be solely relied upon for testing.
- Care should be taken to ensure that virtual data adequately represents the desired complexities of real-world data.
- Not suitable for operational decision making.

4.2.6. Open-source data

Open-source data is publicly available and can be used by anyone.

Considerations for use include:

- Open-source data can supplement existing data sets and feature engineering. For example, the use of public weather datasets may be useful in the performance of environmental impact modelling.
- Data quality, reliability and relevance of open-source data should be validated rigorously, as well as compliance with licensing and data usage policies.
- Adequate use between training, validation and testing will depend on the dataset and need to be assessed.

4.2.7. Adversarial data

Adversarial data is data intentionally used with malicious purposes, either during model training or inference.

Unlike the other data categories, adversarial data does not support model development and is not intentionally used. Rather, it is engineered and introduced by external agents with the malicious purpose to thwart model performance in some way.

TABLE 2. RECOMMENDED USE AND CONSIDERATIONS FOR DIFFERENT POTENTIAL DATA SOURCES

Data Source	Most suited for			Comment
	Training	Validation	Test	
Field	🟩	🟩	🟩	Only suitable type of data for testing
Laboratory	🟩	🟩	🟨	Use for testing needs to be in conjunction with field data
Simulated	🟩	🟩	🟥	While appropriate for early model development, use for testing is not recommended.
Synthetic	🟩	🟩	🟥	
Virtual	🟩	🟩	🟥	
Open Source	🟨	🟨	🟨	Requires thorough review for accuracy, adequacy, etc

🟩	Unlimited/unconstrained use
🟨	Limited/constrained use/requires review
🟥	Not recommended/very limited use

DATA FITNESS FOR USAGE

The overall quality of the data supporting AI/ML applications is crucial. The characteristics that make a dataset suitable, or fit, for use in a given application, and thus of quality, are case-specific. Specific metrics or methods to quantify these attributes will also vary between different applications. While a discussion of such metrics or methods is outside of the scope of this document, it does provide a list of guiding questions that help assess or characterize each attribute as well as clarify its intent.

These considerations apply equally and individually to each of the three subsets of data defined earlier: training, validation and testing data.

4.2.8. Data quantity

Data quantity is the volume of data available for AI/ML applications.

Expected impact on AI/ML models:

- Insufficient data, that is not representative of real-world operations/scenarios, can lead to a risk of overfitting, where the AI model performs well on training data but poorly on new data.
- Having sufficient volume of data enhances the robustness and ability of AI models to generalize.

Considerations for assessment:

- Asses if the volume of data is adequate for training a reliable model for its intended use.

- Consider alternative strategies, such as data augmentation or synthetic data generation, to mitigate data scarcity risk.
- Assess if the data quantity aligns with the complexity of the process or phenomena being modelled.
- Determine if there are any regulatory requirements regarding the minimum amount of data required for validation.

4.2.9. Data relevance

Data relevance refers to the extent to which the dataset is applicable to and describes the desired application scenario.

Expected impact on AI/ML models:

- Data relevance directly affects the adequacy of the models for the intended application.
- Significant impact on performance is expected if model has been trained on data whose characteristics differ from the data at deployment.

Considerations for assessment:

- Determine whether the data accurately reflects the expected conditions at deployment for the intended application.
- Determine whether the data encompasses all expected scenarios during deployment.
- Review which, if any, simplifications or assumptions are inherently included in the data.

4.2.10. Data distribution

The data distribution is the way data points are spread across a range, different categories or classes.

Expected impact on AI/ML models:

- An imbalanced data distribution, or a training/test distribution that differs from the real-world target distribution, can introduce bias into AI models, leading to unfair or inaccurate output.
- Ensuring balanced data distributions is critical to the fairness and accuracy of AI model outcomes.

Considerations for assessment:

- Determine whether the data is representative of the different categories it aims to cover and if the distribution has this been quantitatively evaluated.
- Determine if techniques such as resampling are, or can be, used to balance the data distribution.
- Assess the importance of each class or category in the context of the AI model’s task.
- Determine whether the data distribution reflects real-world conditions or if it is skewed.
- Assess the risk of the real-world distribution changing over time, what impact this would have on model reliability and how this can be monitored.

4.2.11. Data scope

Data scope is the range and context in which the data is applicable.

Expected impact on AI/ML models:

- Data that is too narrow or broad in scope can lead to AI models that are either over-specialized or too generalized for the intended use.
- Properly scoped data ensures that the model and the resources required to train, test, deploy etc are both effective and efficient in their intended applications.

Considerations for assessment:

- Determine whether the data covers all necessary conditions and scenarios for its intended application and whether it covers any conditions and scenarios that are irrelevant for the intended application.
- Consider if there are specific operational contexts where the data scope may need to be adjusted.

4.2.12. Data bias

Data bias is systematic errors in data that can lead to unfair, skewed or poor-quality outcomes.

Expected impact on AI/ML models:

- Data bias can introduce ethical issues, such as discrimination or unfairness, into AI models. For example, an AI used for evaluating resumes and hiring, if trained on a biased dataset has the potential to unfairly favour or disfavour certain groups of applicants.
- Addressing bias is important for the model's reliability as well as maintaining stakeholder trust.

Considerations for assessment:

- Assess if there are known or potential sources of bias in the data.
- Identify what steps are being taken to identify, mitigate or correct data bias.
- Assess whether the data is collected from diverse and representative sources.
- Assess whether there are blind spots in the data that could introduce unintentional bias.

4.2.13. Susceptibility to errors

A datasets susceptibility to errors is the likelihood of inaccuracies or mistakes present in the data and is another measure of data quality.

Expected impact on AI/ML models:

- Undetected errors in training data can propagate through the model, leading to inaccurate or unreliable results.
- Erroneous data can cause misallocation of resources by triggering unnecessary interventions or maintenance activities in the example of predictive maintenance. This highlights the importance of validation of model outputs where possible.

Considerations for assessment:

- Review and identify if there are any data quality and validation checks in place to catch errors in data collection or entry.
- Determine if there is clear ownership and accountability for data sources being used in AI model applications and correction of errors.
- Assess if data lineage is documented for auditability.
- Assess whether there is a system in place for continual monitoring and correction of errors and if there are automated alert mechanisms for identifying and reporting data anomalies or errors.
- Assess the impact of potential errors on the AI model's performance and safety considerations.

4.2.14. Data alignment

A key consideration is the alignment of the data used in model development (training, validation and testing data) with the target application scenario. It is important to note that characteristics such as data relevance, distribution and scope for the testing data define and bind the range of applicability of the model with the performance as demonstrated during testing. No assumptions should be made on the model performance on data that extrapolates from that range. Therefore, it is important to seek alignment between the characteristics of the data used in model development and those expected during deployment of the target application.

4.2.15. Data quality management

As mentioned, the specific required characteristics of a dataset are application dependent. Once those characteristics have been defined, the quality of any given dataset for the desired purpose can be assessed against them.

For data quality management in an AI/ML project, ISO/IEC DIS 5259-1 suggests a data quality management framework for the data life cycle [1]. The data life cycle consists of data requirements, data planning, data acquisition, data preparation, data provision and data decommission as shown in FIG. 3.

This life cycle in ISO/IEC DIS 5259-1 has been modified from ISO/IEC 8183 [2]. The data quality management framework provides the process for determining, accessing and improving the quality of datasets for use in AI/ML applications. The framework includes the following elements:

- Data quality model:
  A defined set of data quality characteristics that provides a framework for specifying data quality requirements and evaluating data quality.
- Data quality measures:
  Means of evaluating each data quality characteristic in the data quality model.
- Data quality assessment:
  Means of assessing whether a dataset meets its needs and requirements.
- Data quality improvement:
  Means of transforming data to improve the dataset’s quality to the extent that it meets the needs and requirements of the organization.
- Data quality reporting:
  Means of publishing data quality reports for determining the root cause of the poor performance of an AI/ML model and for transparency and explainability of the AI/ML.

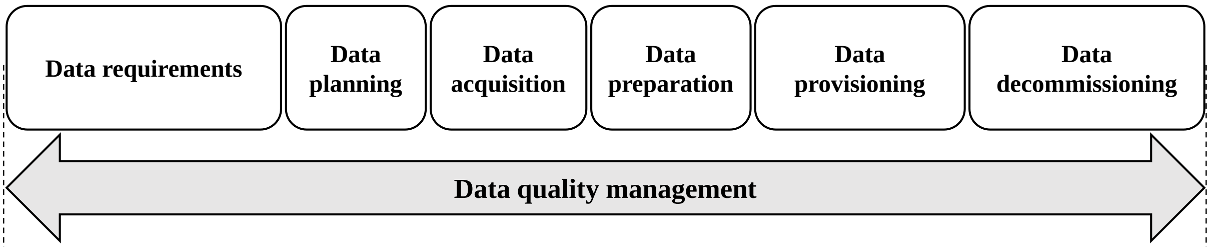


FIG. 3. Data quality management framework in ISO/IEC DIS 5259 [1]

It is relevant to consider the data security of data underlying AI/ML applications and how they may be distinctively affected. The primary goals of data security involve confidentiality, integrity and availability (CIA) and each can be threatened in a different way at different stages of data processing, as indicated in Table 3.

TABLE 3. EXAMPLE THREATS TO DATA SECURITY AT DIFFERENT STAGES OF DATA PROCESSING

	Confidentiality	**Integrity**	**Availability**
Data storage	Disclosure	Tampering	Ransomware
Data transmission	Eavesdropping	Falsification	Transmission interruption
Data processing	Reverse engineering	Data poisoning	Overloaded

Threats to confidentiality and availability to AI/ML supporting data are like those of any other data (such as encryption, authentication and duplication) and thus will not be discussed here. AI/ML applications, however, are subject to specific threats to their data integrity that warrant discussion.

The U.S. National Institute of Science and Technology (NIST) defines data integrity as "the property that data has not been altered in an unauthorized manner" [3]. Data integrity in the case of AI/ML application to NPPs can similarly be defined as the property that data maintain their quality throughout the life cycle of the AI/ML application, not deteriorated by an unauthorized agent or process, where the latter includes inadvertent consequences of changes in operational characteristics of the power plant.

Some of the main threats to data integrity for an AI/ML application include:

- Data drift:
  The unexpected noises contained in the data, or the undocumented changes to the data structure, can cause the AI model to draw inaccurate or incorrect conclusions.
- Data corruption:
  If the data is corrupted such as it is incomplete or unusable, the AI model may produce unstable outputs or fail to work.
- Data poisoning:
  Malicious adversarial data can be injected into the training dataset of AI applications, causing AI model to learn from the poisoned data. AI model will be compromised and make untrusted decision when faced with the new data. This can occur either at the training or inference stages.

Table 4 lists some ways through which each of these threats can present themselves. It is noteworthy that data drift can be inadvertently caused by normal operation of the plant without any intended adverse action. While the causes for the other threats are not specific to AI/ML applications and may already be minimized by existing security and other measures, the causes for data drift require specific measures to be implemented, such as specialized monitoring.

TABLE 4. THREATS TO DATA INTEGRITY IN AI/ML APPLICATIONS

Threat	Presented by
Data drift	• Operational changes in the NPP (even if expected) • Natural changes over time
Data corruption	• Accidental or malicious human activity • Logical errors during the data transfer, processing and storage • Physical compromise to the device hardware or data disk • Sensor malfunction
Data poisoning	• Malware, virus, or cyber attacks • Accidental or malicious human activity

4.3.DATA MANAGEMENT PRACTICES

The management of data integrity is necessary to ensure that data, irrespective of how it was generated or its format, is properly recorded, processed, retained and used in a manner that ensures a complete, consistent and accurate record throughout the data life cycle. This management, called data governance, should be carried out based on the organization's internal standards, policies, rules and processes. The data governance provides the following [1]:

- A set of guiding principles established by an organization to actively manage and improve data quality.
- Decision-making structures and accountabilities through which those assigned with data quality responsibilities are held to account.
- Organizational roles and responsibilities to ensure data quality through repeatable processes.

Some specific considerations for data management include:

- Metadata management
  - Maintenance of metadata ensures data traceability and understandability.
  - Metadata management can play an important role in facilitating data discovery for AI models.
  - Implement practices like a Business Glossary and Data Catalogue for inventory and classification of metadata.
  - Define clear ownership and stewardship across data domains and classes.
- Data lineage tracking
  - Data lineage should be systematically mapped to provide a clear audit trail from source to consumption. This will help in debugging model and data issues and verifying transformations.
  - Record transformations and pre-processing steps applies to the data and ensure data manipulation is properly logged and versioned for traceability.
- Data records
  - Keep a record of all data that impacts AI model decisions, particularly those affecting safety and regulatory compliance. For example, models used for driving condition-based maintenance have training data and model parameters clearly documented in records.
  - Records are stored in secure environments to ensure integrity.
- Data life cycle management
  - Implement data archival and retention guidelines. For example, raw sensor data from a reactor system used in model training and inference may need to be stored indefinitely in order to explain potential failures and gaps in model performance.
  - Regularly backup data to prevent loss.
  - Use secure, encrypted and redundant storage solutions to safeguard data against loss, corruption and unauthorized access.
  - Outline clear procedures for safely decommissioning and deleting data that is no longer needed. Ensure this process meets regulatory requirements and all data lineage information is updated to reflect the decommissioned data.
- Data quality monitoring
  - Implement continuous monitoring systems to identify data quality degradation over time.
  - Utilize quality attribute metrics like clarity, accuracy, consistency and completeness for monitoring.
  - Implement validation rules, cross references with trusted data sources and periodic reviews by subject matter experts as required.
  - Utilize automated alerts for data quality issues including threshold breaches, anomaly detection and other pattern recognition techniques that could indicate quality degradation, corruption or tampering.
  - Quality monitoring and checks should be implemented at different stages of the data lifecycle and automated where possible.
  - Maintain a log of identified data quality issues, along with status and resolutions steps to support issue resolutions and longer-term trending and quality improvement efforts.
  - Establish a feedback mechanism with data stewards and business SMEs to continually improve data quality.
- Need for audits
  - Conduct period audits for compliance with data governance policies and standards across. Audit can be conducted across the full data lifecycle and/or AI models.
  - Audit logs and reports should be stored securely and results trended to support future investigation and program performance analysis.
- Data leakage
  - Data partitioning – It is important to carefully partition data into training, validation and testing sets while ensuring no data overlap between them. Consider sampling techniques that maintain the distribution of classes across different sets.

4.4. DATA SHARING

Data is fundamental to enable successful AI/ML solutions. Furthermore, field data is arguably the most important of the data sources discussed as it is the only source that enables relevant and meaningful test and assessment of the application. In the case of NPPs, this data resides with the nuclear operators, who often lack the expertise and resources to develop AI/ML solutions on their own. This scenario makes data sharing a crucial enabler for the successful deployment of AI solutions across the nuclear industry. In the context of this document, data sharing refers to nuclear operators, as the data owners, sharing relevant field data with AI/ML solution developers, or other operators in a combined effort, or with the AI/ML community at large to support the successful development and deployment of AI/ML solutions of interest.

Despite its importance, it has been difficult to achieve a culture and practice of data sharing within the nuclear industry to the level necessary to support the successful development and deployment of AI/ML tools at large. However, some success is starting to be experience in some isolated cases. The reasons for the reluctance in sharing data are understandable, as often the relevant data is sensitive and sharing it can lead to adverse impacts to the data owner for diverse reasons. Nonetheless, the benefits of sharing could overcome the associated risk, especially if the necessary care is taken. The following identifies the value of sharing data and some of the practices to overcome associated concerns.

4.4.1. Value and importance

Some of the benefits brought about by data sharing include:

- Fostering a culture of collective problem-solving and innovation, leading to more robust AI models and analytics tools. It is often the case where nuclear operators are rich in real world data but lack the specialized skills and tooling to perform advanced research and implementation of AI models using this data. On the contrary, academia and private firms that have these specialized resources often do not have access to this rich real-world data. Collaboration between these types of parties therefore supports advancement of innovation as well as results for the nuclear power operators.
- Helping in the formation of industry-wide standards and best practices, which is particularly important in the regulated nuclear energy sector.
- While not specifically an AI related benefit, it can also support regulatory compliance and improving efficiency of regulator audits through reduction in resources and effort required during discovery and data collection. This is an important consideration for activities that require regulatory oversight: although not necessarily (or initially) required, it is on the operator's best interest to share as much data and related information as possible with the regulators; this will not only facilitate and expedite the review by minimizing additional requests but also help foster trust and transparency.
- For some applications of common interest across the nuclear fleet, it is not expected that any one single operator would have sufficient relevant data to support successful development and deployment of AI/ML tools, so data sharing in a collaborative effort is the only feasible way to realize the envisioned benefit.
- Data representativeness has been identified as one common key characteristic of the datasets for successful deployment of AI/ML solutions. Through data sharing, operators can guarantee that their assets are represented in the underlying data feeding AI/ML models, thus increasing the likelihood for successful model deployment in their plants.
- Creation of relevant benchmark sets that enable the industry to assess the performance of proposed solutions on meaningful data in a safe environment and before committing significant resources for deployment. Such datasets would allow the industry to easily evaluate solutions from different providers against common, well-understood and documented scenarios. It also allows solution providers to assess their solution against relevant data that is otherwise typically unavailable to them.

4.4.2. Confidentiality and sensitivity concerns

Some practices that can aid in addressing or minimizing data sensitivity concerns include:

- Confidentiality concerns can often be addressed by anonymizing or pseudonymizing data prior to sharing.
- Data masking techniques can also be used to hide specific sensitive attributes.
- Export licenses may also be pursued through regulators to share data as required by local laws, regulations and operating licenses.
- Utilize data sharing agreements between parties that specify the terms of use, rights and responsibilities etc.
- Ensuring data sharing respects ethical considerations like individual privacy.
- Adhering to data sovereignty laws.
- Securing data for sharing from assets that are no longer in operation (e.g., decommissioned plants or retired components) while the data is still accessible. Help identify the existence of such data.
- Supporting activities where the data can be accessed in a safe and controlled manner. One example is field trials, where data can be accessed locally and be used to inform and test models without having to leave the site.

4.5.MANAGING ADVERSARIAL USE/ CYBERSECURITY

The adoption of digital technologies to support the operation and maintenance of nuclear facilities is expected to have a wide range of benefits for optimum control, improved operational flexibility, predictive maintenance and better inference of uncertainties, etc. Combining digitization with networking and high-fidelity simulations, collectively referred to as a cyber physical system (CPS), the system self-awareness can be greatly improved, allowing for improved diagnostics, prognostics and optimal response to various process changes. However, this wave of digitization has heightened concerns about the vulnerability of digitization to intrusive attacks, often referred to as the cyber vulnerability or cybersecurity problem. The goal is to assess the level of control an adversarial entity could have if it establishes a foothold in the digital network of a CPS for near- or long-term malicious purposes, e.g., denial or interruption of service, degraded performance, etc. The increased frequency and sophistication of recent cyberattacks against CPSs makes it clear that reliable defences must be established to ensure zero impact on the system function if its digital network is compromised.

The initial response to these threats has focused on the adoption of information technology (IT) defences, aiming to stop unauthorized access via firewalls, passcodes, VPNs, etc. With the ever increase in sophistication of cyber intrusion methods, recent attacks have proven that IT defences can be eventually bypassed by persistent adversaries, e.g., state-sponsored or criminal organizations. Among the most famous attacks is the state sponsored Stuxnet malware which compromised the control network of the Natanz Iranian enrichment plant leading to its shutdown after damaging thousands of its centrifuges [25]. Other similar daring attacks occurred thereafter, e.g., attacks against the Japanese Monju nuclear power plant [26], the Gundremmingen nuclear power plant in Germany [27], the Electric Grid in Ukraine [28] and the most recent one happening in India [29].

In light of these threats, it has become essential to build another layer of defense when IT defences are compromised. This new layer is referred to as the operational technology (OT) defense or the physical process defense. The OT defences focus on the physical process as described by the network data comprised of sensors readings, process variables and actuating commands as well as the physics models that are often interfaced with the network. OT defences ask the question: are the engineering network data consistent with expected behaviour? In a sophisticated false data injection (FDI) attack, the attacker relies first on delivering an IT payload, designed to penetrate through the IT defences, representing the conventional first step for any hackingh attempt, i.e., gaining access to the system. Following that, the attacker must deliver another payload, referred to as the engineering payload. This payload is designed to cause the system to move along an undesirable trajectory [30]. This can be achieved in multiple manners. For example, the engineering payload could falsify the sensors data, causing the control algorithms to send signals to the actuators that cause the system to exceed its safety margins. Another approach is to change the control algorithm logic to achieve similar goals. In all scenarios, the payload must be aware of the normal engineering checks that exist in the network. These checks are developed by the engineering team to ensure that system is reliably responding to normal process variations. Thus, unlike IT defences which rely on the use of generic methods to protect access to information, OT defences must be cognizant of the engineering design and safety procedures in place. To achieve that, OT defences must rely on an online monitoring approach to continuously check the engineering data, i.e., sensors readings, process variables, etc., and be able to determine whether the data are genuine.

This represents a great potential for employing AI/ML techniques, which proved extremely powerful in detecting trends in complex datasets. There are two general approaches to achieve automation of pattern recognition for monitoring purposes: passive and active. Passive monitoring implies observing the system for a time period to understand, i.e., establish a basis to describe, its normal behaviour and use this understanding to judge whether the system behaviour at a later time has deviated from its expected normal behaviour [32]. Active monitoring relies on injecting signature perturbations into the engineering data to ensure their trustworthiness [32], [33]. Effectively, active monitoring may be thought of as a data fingerprinting methodology for authentication. The logic behind passive monitoring is that one needs to observe to develop an understanding, while active monitoring argues that one needs to interfere to develop an even better understanding.

Both viewpoints have their merits and challenges. For example, a key challenge of passive monitoring is that the attacker is expected to do the same thing done by the defender, i.e., observe the network during a lie-in-wait period before introducing changes to the data. The more familiarity the attacker has with the system the easier it is to build an understanding of its behaviour, ultimately reaching the level of understanding of the OT defense designer. In active monitoring, the first challenge is to ensure that the injected perturbations do not penalize the system function. Second, if the embedded fingerprints follow a pattern that can be detected using passive monitoring, the whole purpose of the active monitoring is defeated, because the attacker can learn their patterns and ultimately circumvent them. These challenges require good understanding of system behaviour, achieved via passive monitoring.

5. IMPLEMENTATION CONSIDERATIONS

5.1. PRE-IMPLEMENTATION CONSIDERATIONS AND POTENTIAL USES

AI can play a role in the emergent failure diagnosis, troubleshooting and corrective actions. Instances have occurred of emergent failures of equipment that require immediate action to restore a needed function or indication. These failures may require expertise to diagnose, troubleshoot and repair. AI can provide the ability to promptly diagnose, troubleshoot and repair without delays with contacting maintenance during off hours and or weekends. To achieve this AI application utilities will need to build a dynamic learning model of troubleshooting and repair strategies from historical database and continue to use available OPEX to augment the AI recommendations. The feedback loop may require additional steps in the engineering and maintenance process of work planning and work order documentation. The dynamic feedback will allow the AI model to continue to be refined with newly identified failure modes, improved maintenance methodology and improved predictive equipment monitoring of degradation with associated failure precursors.

To achieve this, AI project dedicated hardware and software solutions will need to be developed, deployed and maintained. The integrated hardware and software solution will require management software and hardware storage for raw input data to be used in the AI algorithm. In addition, the AI algorithm will need to be executed on a periodic frequency to achieve the dynamic learning model update. The raw data content will need to be managed and stored for access to the AI. This raw data may include text, video and other related content. The source of the data for this application will be the equipment performance trending to achieve improved degradation precursors. The trending data may include factors such as pre-failure equipment temperature, pressure, vibration, oil analysis and inspections data collected prior to the equipment failure. This data may also include equipment specific maintenance performed, troubleshooting performed and previous successful repair strategies. This data will be collected as part of the maintenance execution and documentation process.

AI can play a role in assisting operators to respond to complicated multiple casualty transients by providing insight into the priority procedure to implement based upon risk and historical data. If an event has competing applicable procedures that need to be implemented, we currently depend upon Operations leadership to prioritize mitigation strategies based upon limited experience and assessment. AI can determine the best option amongst competing strategies including abnormal procedures, emergency procedures, fire response, flood response, emergency procedures, flex procedures and/or severe accident procedures.

To achieve this AI application utilities will need to build a dynamic learning model of transients and simulated transients from historical databases and continue to use current OPEX to augment the AI recommendations. The feedback loop will require additional steps in the operations/engineering transient review from actual and simulated plant data. The dynamic feedback will allow the AI model to continue to be refined with newly identified failure modes, improved operation accident mitigation and improved procedure prioritization during complex abnormal or emergency operations.

AI can be applied to various applications in the nuclear power industry due to its ability to automate tasks, analyse complex data and satisfy various reporting and regulatory requirements. It can also be used to automate the collection of data from various sensors and monitoring systems in a nuclear power plant. This ensures that data is continuously gathered and reduces the need for manual data collection, which can be time-consuming and prone to errors.

AI can analyse vast amounts of data quickly and accurately. In the nuclear industry, this means AI can detect anomalies or deviations from normal operating conditions in real-time. It can also predict equipment failures or maintenance needs, helping plant operators take proactive measures.

AI can generate detailed reports and insights from the data it collects and analyses. These reports can be used for regulatory compliance, performance monitoring and decision-making.

5.2.REVIEW OF EXISTING REGULATORY REQUIREMENTS

Currently, regulatory requirements around the use of AI in nuclear related activities are nascent. However, national nuclear regulators are taking steps to move towards providing at least guidance for those who seek to deploy AI in the nuclear realm. Two such examples will serve to illustrate those efforts.

5.2.1. Canadian Nuclear Safety Commission – United Kingdom’s Office of Nuclear Regulation and the United States Nuclear Regulatory Commission (CANUKUS) White Paper on AI

The Canadian Nuclear Safety Commission (CNSC), the United Kingdom’s Office of Nuclear Regulation (UKONR) and the United States Nuclear Safety Commission (USNRC) have a long history of collaboration on regulatory matters. While each regulator has undertaken work separately to understand and characterize the potential safety benefits and risks of deploying AI in nuclear activities, in 2023 they began a tri-lateral cooperative project to publish a tri-later White Paper or Principles Paper on this topic.

This paper is scheduled to be published simultaneously on all three regulator’s external websites in June of 2023. It is expected to describe an approach to the regulation of AI in nuclear activities and with regards to nuclear material materials in Canada, the United Kingdom, and the United States. The paper promises to explain how common objectives and consideration of areas important to the effective regulation of AI across all three countries is possible. It is intended that the considerations outlined will encourage beneficial uses of AI and clarify the challenges arising from these fast-developing technologies and the principles applied to regulating them.

The paper will touch on the following:

- AI Use Cases – High Level Categories;
- Country-Specific Regulatory Frameworks;
- Use of existing safety and security systems engineering principles;
- Human and Organisational Factors;
- Architecture;
- Lifecycle Management;
- Demonstrating adequately safe and secure systems that contain AI.

5.2.2. United Kingdom’s Office of Nuclear Regulation and Environment Agency Report: Pilot of a regulatory sandbox on artificial intelligence in the nuclear sector

Nuclear regulators are employing innovative approaches to conduct preliminary assessment of how best to allow for the deployment of innovation itself into nuclear regulated activities. One such example is the use of “Sandboxing” by the ONR. In coordination with the UK’s Environment Agency the ONR.

The sandboxing process involved sprint workshops to consider the key aspects associated with the deployment of AI in the two problem/opportunity statements and the associated mock safety, security and environment case structures. These key aspects were then prioritised into four deep dive topics for each of the two AI applications and explored through regulatory sandboxing sessions.[3]

The report on the effort stated that benefits of AI should be clearly articulated, and the how the benefit compared to existing technologies should be taken into consideration. The risks associated with the deployment of AI need to be characterized, understood and mitigated if need be. The report also made the following recommendations:

- Deploy AI gradually and in a phased manner to build confidence and experience;
- Evaluate whether a principles-based approach to regulation is preferred, to take into account differing considerations for each potential application of AI;
- Understand the limitations of training data, especially as it applies to deployment with real data;
- Understand the importance of human factors in the AI life cycle, and;
- Establish an AI safety culture to complement and support conventional safety culture.

5.3.PRE-OPERATIONAL DEPLOYMENT CONSIDERATIONS

5.3.1. Level of human oversight or autonomy

Chapter 1 discussed representative levels of human oversite for an AI system, or, alternatively, how much autonomy the AI has on controlling the underlying system. These levels range from an AI providing insights into data collected on the system for a human analyst to full autonomous control of the system with little to no human oversite. The level of human oversite should be considered when deploying an AI system. The role of the human in the functioning of the system should be defined, as well as how the human should interact with the system, along with the appropriate level of training needed to interact with the AI. This could also influence the implementation strategy if the AI if more control is granted to the system as more experience is gained with the system. The level of human oversite on the system could play a role in the threshold of acceptance for the system, but this may not be only factor. The algorithm selection, and overall risk of the application, may also impact the final acceptance of the application. The overall risk of the system encompasses both the risk level of the application, but also how the AI is deployed.

5.3.1.1. Human-in-the-loop (HITL) systems

Some AI systems operate with a high degree of oversight, often referred to as Human-in-the-Loop (HITL) systems. In these cases, humans are actively involved in the decision-making process. These systems are used in contexts where human judgment and intervention are crucial, such as medical diagnosis, autonomous vehicles, or military applications. HITL systems are considered safer and more accountable but they may be slower and more expensive.

5.3.1.2. Human-on-the-loop systems

[3] Outcomes of nuclear AI regulatory sandbox pilot published - Office for Nuclear Regulation - News (onr.org.uk)

In these systems, humans are involved in monitoring and supervising AI but they are not actively making decisions during operation. Human-on-the-Loop systems provide a level of oversight while allowing AI to operate autonomously. Examples include AI-driven content moderation on social media platforms or automated trading algorithms in finance.

5.3.1.3. Human-out-of-the-loop systems

On the opposite end of the spectrum, there are AI systems that operate with minimal or no human intervention. These systems are designed to be fully autonomous and are often used in applications where rapid decision-making and scale are critical, such as recommendation systems or high-frequency trading algorithms. However, these systems can raise concerns about accountability, bias and safety.

5.3.1.4. Ethical and regulatory considerations

The level of oversight and autonomy should be determined by potential impact to safety, risks ethical and regulatory considerations. For example, in fields like healthcare and autonomous vehicles, there are strict regulations and ethical guidelines that mandate a high level of human oversight. In contrast, in industries like e-commerce or online advertising, the autonomy of AI systems is often higher.

5.3.1.5. Balancing autonomy and oversight

Striking the right balance between autonomy and oversight is crucial. Too much autonomy may lead to unintended consequences, while too much oversight can hinder the potential benefits of AI. The level of oversight should be adjusted to suit the specific goals and risks of a given application.

5.3.1.6. Transparency and accountability

Regardless of the level of autonomy, transparency and accountability mechanisms are essential. This includes tracking and explaining the decisions made by AI systems, providing avenues for recourse when errors occur and conducting audits to ensure compliance with ethical and legal standards.

5.3.1.7. AI development stages

AI systems may evolve through different stages of autonomy during development. Starting with supervised learning and gradually moving towards reinforcement learning can help ensure that the AI system is properly trained and monitored as it becomes more autonomous.

5.3.1.8. Continuous monitoring and adaptation

AI systems should be continuously monitored and adapted to changing circumstances. This can involve retraining models, updating algorithms and refining the level of human oversight as the technology matures.

The level of oversight or autonomy in AI systems is a nuanced and context-dependent decision that should consider ethical, legal and practical factors. Striking the right balance is essential to harness the benefits of AI while minimizing risks and ensuring accountability.

5.3.2. Training metrics

The developer of an AI solution should outline and discuss the training metrics used to evaluate the performance of the AI model. Section 3.3.8 provides further discussion on potential training metrics applicable for AI systems. Understanding and choosing the right evaluation metrics is critical in AI model development. The choice of metrics should align with the specific goals, use case and potential consequences of the model's predictions, however it is important to consider the trade-offs between accuracy, false positives and false negatives. Balancing accuracy, false positives and false negatives depends on the specific use case and the associated costs or consequences of each type of error. Sometimes, optimizing for one may come at the expense of another. Training a model to the highest accuracy, or quoting the accuracy of the trained model, may not be sufficient to support model development. For example:

- AI models applied to biased datasets may make interpretation of an accuracy metric challenging.
- If the consequences of false positives are unacceptable, a developer may need to sacrifice model accuracy and precision may be the desired training metric.
- If the consequences of false negatives are unacceptable, a developer may need to sacrifice model accuracy and recall may be the desired training metric.
- Selection of Mean Absolute Error, Mean Square Error, Root Mean Square Error, etc.

The methods used to evaluate the training metric is also important and should be defined. If a train/test split is created for development, or a holdout set is retained for testing, what data is used to calculate the reported training metric provides context on what the training metric means.

5.3.3. Initial and continuing operator training AI

For systems where humans are integral to the operation of the AI, the performance of the user should also be considered when evaluating the performance of the system. This is important when a user is providing oversite of an AI system and when an AI is providing recommendations to a user. For instance, if a human is in the loop to provide a backstop for the performance of the AI system but the human accepts all recommendations from the AI, the human is not serving its function as a backstop. This behaviour would influence the “accuracy” of the deployed system. The operator of the system should understand both the performance of the underlying physical system and the AI application. The role of the human for the operation of the system should also be defined. The operator should receive the appropriate level of training to use/monitor the AI application but also understand the underlying physical system to be able to judge when the AI is functioning appropriately. Therefore, the operator would understand what they should be monitoring and when they should provide oversite, as well as what they should do if oversite is required.

5.4.ENGAGEMENT AND IMPLEMENTATION STRATEGY

5.4.1. Identification and documentation of risk

AI can play a crucial role in enhancing safety and operational efficiency at nuclear power plants. To address the identification and documentation of risk in the context of AI at nuclear power plants, it is essential to consider potential issues, vulnerabilities, behaviour attributes, consequences of failure and AI response.

5.4.1.1.What could go wrong?

- **Vulnerabilities:** Vulnerabilities in AI systems at NPPs may arise from various sources, such as errors in data input, software bugs, or external cyberattacks. These vulnerabilities can compromise the integrity, reliability and safety of AI systems.
- **Behaviour attributes**: Understanding the behaviour attributes of AI is vital. AI systems can exhibit unexpected behaviours due to data anomalies, model drift, or adversarial attacks. It is important to monitor and validate AI systems continuously to detect any abnormal behaviour.
- **How do vulnerabilities manifest:** Vulnerabilities in AI may manifest as incorrect predictions, misclassifications, or other deviations from expected behaviour. For example, a vulnerability in a predictive maintenance AI system might lead to a missed detection of a critical equipment failure**.**

5.4.1.2.Consequences of failure

- **What happens in case of failure of the application:** The consequences of AI failure in a nuclear power plant can be severe. AI systems are often used for tasks like equipment health monitoring and anomaly detection. If an AI system fails to perform its duties correctly, it can lead to safety incidents, equipment failures, or disruptions in plant operations.

- **How does AI respond to vulnerabilities:** AI systems must be designed with robustness and fault tolerance in mind. They should include mechanisms for self-assessment, self-correction and adaptation. For example, if an AI system detects that its performance is degrading or that vulnerabilities are present, it should trigger an alert, notify operators and potentially switch to a safe mode of operation. Additionally, AI systems can be equipped with cybersecurity measures to defend against external threats. AI should be designed to supply reliable exit mechanisms for users should they be needed to alleviate, avoid or escape from the results of a vulnerability.

To mitigate these risks and consequences, nuclear power plant operators and AI developers can take several measures:

- **Comprehensive testing:** Thoroughly test AI systems under various conditions, including edge cases, to identify vulnerabilities and ensure that they behave as expected.
- **Redundancy:** Implement redundancy in critical AI systems to ensure that there are backup mechanisms in place if a primary system fails.
- **Regular auditing and monitoring:** Continuously audit and monitor the AI system's performance and behaviour, looking for signs of vulnerabilities or deviations from expected behaviour.
- **Cybersecurity measures**: Strengthen cybersecurity to protect AI systems from external threats and ensure the integrity and confidentiality of data.
- **Training and education:** Train and educate personnel on AI systems, their limitations and how to respond to potential issues.
- **Emergency response plans:** Develop comprehensive emergency response plans that outline procedures for handling AI failures or vulnerabilities to minimize their impact on plant operations and safety.

The use of AI at nuclear power plants can provide significant benefits but it also brings the need for robust risk assessment, monitoring and mitigation strategies. Identifying vulnerabilities and understanding AI's behaviour attributes are essential for ensuring the safe and reliable operation of AI systems in this critical environment.

5.4.1.3. How to mitigate or preclude negative consequences?

To mitigate or preclude negative consequences in the context of AI at nuclear power plants, it is crucial to implement safeguards and contingency plans. This includes having guardrails to limit AI actions and employing defense-in-depth strategies.

5.4.1.4. Does application need guardrails to limit use?

- **Constraints and thresholds:** Implement constraints and thresholds in the AI system to prevent it from making decisions or taking actions that fall outside safe operational boundaries. For instance, an AI system for reactor control should be constrained by strict limits to prevent it from making decisions that could lead to a meltdown.
- **Human oversight:** Maintain human oversight and control over critical decisions. While AI can provide recommendations and automated processes, there should be human operators who can intervene and override AI decisions if necessary. This human-AI collaboration can act as an additional layer of protection.
- **Auditing and validation**: Regularly audit and validate the AI's behaviour against predefined safety rules and guidelines. If the AI system starts to behave in an unexpected manner or approaches predefined boundaries, it should trigger an alert for human intervention.

5.4.1.5. Is defense in depth available if system fails?

- **Redundancy:** Implement redundancy not only in the AI systems but also in the entire control and safety infrastructure. This means having backup systems that can take over in case of AI system failure. For example, if an AI system used for emergency shutdown fails, there should be a redundant manual or automated backup system in place.

- **Isolation and containment:** Ensure that AI systems are isolated from critical safety-critical systems, such as reactor control. This separation prevents a failure in the AI system from directly affecting the core operation of the plant.
- **Emergency response plans:** Develop comprehensive emergency response plans that outline step-by-step procedures for addressing AI system failures or vulnerabilities. This includes clear roles and responsibilities for human operators and established protocols to restore safe operation.
- **Cybersecurity measures:** Robust cybersecurity measures should be in place to protect AI systems from external threats. Regularly update and patch software, use firewalls and employ intrusion detection systems to detect and respond to potential attacks.
- **Training and drills:** Regular training and simulation drills should be conducted to ensure that plant personnel are well-prepared to respond to AI system failures or other emergencies. This includes understanding the procedures, knowing how to switch to manual control and managing communication during crises**.**

Mitigating or precluding negative consequences in the context of AI at nuclear power plants involves implementing guardrails to limit AI actions, ensuring human oversight and having a defense-in-depth strategy in place. Defense-in-depth includes redundancy, isolation, emergency response plans, cybersecurity measures and training to handle AI system failures effectively while maintaining the safety and integrity of nuclear plant operations.

5.5.GRADED APPROACH AND RISK-INFORMED REGULATORY APPROACHES

A graded approach or risk-informed regulation involves tailoring regulatory requirements based on the level of risk associated with specific activities or systems. When applied to the oversight of AI in the nuclear industry, it implies a systematic and flexible approach to ensure safety and security while accommodating advancements in technology. Elements the regulator may consider when applying a graded or risk-informed regulatory approach may include the following:

- **Risk assessment:**
  - Identification of risks**:** Conducting a thorough risk assessment to identify potential risks associated with the use of AI in nuclear applications. This includes understanding the consequences of AI failure, potential vulnerabilities, and the impact on safety and security.
  - Categorization of systems: Classifying AI systems based on their significance to safety and security. High-risk systems may include those involved in critical decision-making or control processes, while lower-risk systems may have less direct impact on safety.
- **Graded approach:**
  - Tailoring regulatory requirements: The regulator may apply a graded approach by tailoring regulatory requirements based on the assessed risk levels. High-risk AI applications may be subject to more stringent regulations, while lower-risk applications may have more flexible requirements.
  - Scalable oversight: The regulator may implement a regulatory framework that allows for scalable oversight. Critical AI applications may undergo more extensive regulatory scrutiny, while less critical applications may have streamlined regulatory processes.
- **Performance-based regulation:**
  - Focus on outcomes: Some regulators may adopt a performance-based regulation approach, at emphasizing achieving safety and security outcomes rather than prescribing specific technology or methods. This allows for flexibility in incorporating new technologies like AI while maintaining a focus on overall risk reduction.
  - Continuous improvement: Regulators are also likely to encourage continuous improvement in safety and security practices by setting performance goals that can be adapted as technology and industry practices evolve.
- **Regulatory engagement:**
  - Collaboration with stakeholders: Regulators are actively engaging with industry stakeholders, including AI developers, operators, and end-users, to gather insights into AI

    - technologies and applications. Collaborative efforts are helping regulators stay informed about the latest developments and ensure that regulatory approaches towards AI remain relevant and fit for purpose.
    - Feedback mechanisms: Regulators are establishing feedback mechanisms to receive input from industry participants regarding the effectiveness of regulatory regimes and this dialogue assists regulators stay agile and responsive to emerging challenges and opportunities posed by AI.
- **Training and competence:**
    - Human factors consideration: Recognize the importance of human factors in AI integration. Ensure that personnel responsible for AI systems are adequately trained and competent to understand, operate, and address potential issues associated with AI technologies.
    - Competency assessments: Develop mechanisms for assessing the competence of personnel involved in AI-related activities. This may include training programs, certification processes, and ongoing competency assessments to ensure a high level of expertise.
- **Periodic safety reviews:**
    - Periodic assessments: Regulators may request periodic safety reviews and assessments to re-evaluate the risk associated with AI technologies, keeping the use of AI aligned with the evolving state of technology and industry best practices.
- **Security and ethical considerations:**
    - Cybersecurity requirements: Regulators may seek evidence of integrate of cybersecurity requirements into the regulatory framework to address potential vulnerabilities associated with AI systems.
    - Ethical standards: Incorporating ethical considerations into the AI framework, emphasizing responsible AI development and deployment may include addressing issues such as bias, transparency, and accountability in the use of AI in decision-making processes.
- **Public communication:**
    - Transparent communication: Regulatory bodies may seek evidence of transparent communication with the public regarding AI in the nuclear industry. It is possible the regulator may wish to communicate regulatory decisions, safety assessments, and risk mitigation strategies relevant to AI to build public trust and confidence in the use of AI technologies in the nuclear industry.
    - Public input: Regulators may seek to encourage public input in the regulatory process, particularly when it comes to high-impact AI applications. Soliciting public perspectives may assist regulators consider a diverse range of views and concerns, contributing to a more comprehensive regulatory framework.

By integrating these elements into a graded approach or risk-informed regulatory framework, nuclear regulators can effectively oversee the use of AI in the industry, balancing innovation with safety and security requirements. This approach enables the responsible integration of AI technologies while minimizing risks and ensuring the continued safe operation of nuclear facilities.

5.6.EXPLAINABILITY

Explainability is a critical factor when it comes to using AI at nuclear power plants as the techniques provide information to the user on why the system performs the chosen action. Different AI techniques have different levels of explainability and it is important for the developer to determine if the technique chosen for an AI solution has an appropriate level of explainability commensurate for the application.

Trade-offs exist between model selection and model explainability. Black box models, such as deep neural networks, are highly accurate but challenging to interpret. They make decisions based on complex, nonlinear relationships within the data, which are not readily explainable. In situations where transparency is critical, like in nuclear plants, black box models may be less suitable. Explainable models, like decision trees or linear regression, provide human-understandable insights into why a particular decision was made. These models are interpretable and can provide insights into the key features and factors driving their predictions.

There is often a trade-off between accuracy and explainability. Highly accurate models, especially deep learning models, tend to be more complex and less interpretable. Achieving a high degree of explainability might result in a sacrifice in predictive accuracy. Finding the right balance is essential. In safety-critical applications like nuclear power plants, explainability is usually prioritized over raw predictive accuracy.

It's essential to balance the accuracy of AI models with their explainability. The model developer should identify the appropriate modelling framework and what advantageous features that framework provides relative to the level of explainability achievable.

How the techniques work and what the output of those techniques imply about the performance of the AI systems should be understood when using them to make a safety claim about the performance of the system. Deliberate choices between accuracy and explainability, selecting suitable model types, ensuring alignment with the task and employing various methods for understanding and interpreting AI decisions are important factors when considering the development and deployment of an AI system. The specific choice of methods should align with the critical requirements of the nuclear plant's operation and safety.

5.6.1. Goal alignment

Alignment refers to the degree to which the AI's training objective matches the actual task it is intended to perform. In the context of nuclear power plants, alignment is crucial. The AI model should be trained with a clear understanding of the specific objectives and requirements of the plant, as well as how to translate the operating objectives into an appropriate training method. Misalignment can lead to incorrect or unsafe decisions or unexpected behaviour.

Ensuring alignment involves careful selection of training data, loss functions, objective function and performance metrics to reflect the task's real-world goals and safety considerations. Explainability techniques can help ascertain if the system is making the decision for the appropriate reasons but may not be completely sufficient. Effort should be undertaken to ensure the objective of the deployed system and the objective of the training, are in alignment.

5.7.VALIDATION

5.7.1. AI specific implications

AI training methods can be trained using a training set along with an accompanying validation set. The validation set is intended to reflect data the system does not encounter during the training phase and presents how well the model performs on new data. Typically, this validation set is sampled from the data available and should have a similar distribution to the underlying dataset. While this process provides an indication of model performance to novel data, it is not fully sufficient to validate the model because model training is dependent on data and it is not possible for the data to be reflective of all conditions the model could encounter.

The nuclear industry has performed significant research on validation, verification and uncertain quantification of computer algorithms. These traditional techniques may, or may not, be robust enough to enable the validation of AI-enabled systems. The scientific and industrial community is researching how to perform V&V on AI-driven systems and what is the degree of certainty of the performed analyses. The use of traditional V&V techniques fails in many cases, when applied to AI-driven systems. AI-specific V&V challenges may manifest if perturbations are introduced to the systems following retraining or shifts in training data, such as: lack of determinism (i.e., given the same input, the system produce the same output in the same execution time), measures of parameters such as accuracy and response time and validate and demonstrate all provided measures in compliance with the requirement of nuclear safety standards. Therefore, how perform V&V of AI-driven systems is currently an opened and capital research question.

Validation techniques can also indicate limits of operability for the AI system. For example, if validation indicates the AI system is not robust over a certain range of operation, restrictions can be placed on the system to inhibit operation, or provide an indication to operators to take control,

There are two main directions in which AI could be deployed in the development and operation of nuclear systems. One is the embedding of AI into operational system components, where given the input from a range of heterogeneous sensors, the embedded AI components calculates and returns inferences on what the system should do next, in the subsequent iteration of control. The use of this type of system raises at least two challenges: how to ensure the quality of the sensor data and how to correct them in case of errors or interferences, e.g., due to noise.

The second direction concerns the use of AI for the V&V of systems, on the verification side of the process, outside of the system itself. In this direction, generative-AI systems based on domain-specific specialization of large language models, (a-la Chat-GPT) could be used to provide procedural advice: safety engineers and verification experts would ask the generative-AI engine questions on how to perform specific V&V procedures on the system under examination and the generative-AI engine would return operational suggestions, such as for example, code fragments, or improvements over solutions proposed in the query itself.

5.7.2. Use of real-time plant data to confirm or refute effective strategy execution

AI can provide the operator with real time plant performance data to either confirm the proper execution of mitigation strategies or identify the strategy is not being effective as executed. This will allow the operator to identify that a current strategy is not working or is not effective. Therefore, allowing the operations crew to select alternative strategies of a defense-in-depth option.

To achieve this AI application utilities will need to build a dynamic learning model of simulated and actual event analysis from historical databases, industry OPEX and current real time data of operational experience to augment the AI recommendations. The feedback loop will require additional steps in the analysis of operations performance during actual or simulated training events. The dynamic feedback will allow the AI model to continue to be refined with improved operational strategies and methodology and improved operator performance during abnormal and emergency events.

5.7.3. Operator actions, HRA/PRA calculations and improvements AI

Operator error rate and equipment reliability data in PRA/HRA models can be improved using augmented AI to recommend and check for proper implementation of mitigating strategies in the abnormal and emergency operating procedures. The current methodology for PRA modelling of operator actions and the actual operator associated underlying performance can be enhanced using AI to confirm proper operator actions and implementation of proceduralized accident mitigation strategies. AI plays a role is ensuring the operator does not make an error or identify an error occurred so the operations crew can correct the error and ultimately correctly execute accident mitigation procedures. For a given set of human-defined objective, AI can make predictions, recommendations, or decisions influencing real or virtual environments. This will improve the HRA/PRA credit given to operator actions and also improve actual operator performance. Use of AI technologies can improve operational performance and mitigate operational risk. This will also be reflected in a more accurate PRA model using AI generated reliability data for human error rates and equipment error rates.

5.8.REGULATORY ENGAGEMENT

Engaging with regulators when applying AI in nuclear power plants is a crucial step to ensure compliance with safety and operational standards. It is essential to approach regulators in a timely manner and address key considerations. Considerations regarding whether the regulator should be engaged regarding the uses of an AI system include:

- **Does the application affect the licensing basis?**
  It is important to approach the regulator, even if the use of AI is determined to not affect the licensing basis. As proactive approach helps ensure transparency and regulatory awareness. While it may not result in a change to the licensing basis, it could lead to discussions or guidelines that clarify regulatory expectations for AI applications.
  A use of AI, even if not affecting the licensing basis, could potentially result in negative inspection findings if it leads to deviations from established safety procedures or standards. Therefore, even in non-licensing-basis-impacting applications, communication with the regulator is essential to preclude adverse inspection findings.
- **Does the application affect the safety of unit?**

Should the AI application have the potential to affect the safety of the unit, it is imperative to engage with the regulator. Any changes that impact safety must undergo rigorous regulatory review and approval.

- **Does the application affect the safety classification of existing systems?**
  If an AI system is judged to have an impact on the safety of an NPP, the underlying supporting infrastructure may require a safety assessment to understand how the AI, with safety significance, may function if a supporting system fails. For example, if an existing sensor is used as input to an AI system which is judged to be safety significant, the sensor should be examined to understand how its performance effects the system. If the performance of this sensor is judged to be important to the functioning of the AI system, the safety classification of that sensor should be assessed because the usage of that sensor's output may have changed relative to how it was traditionally utilized.

Early engagement can let the regulator know your plans, but when approaching a regulator on the deployment of an AI application, it would be advantageous to describe the implementation strategy of the system. Consideration of the implementation strategy early in the development process can assist in a fieldable system. Topics of consideration can include:

- Supporting infrastructure: Describing the infrastructure supporting the AI application, including hardware, software and cybersecurity measures ensures that the supporting infrastructure is robust, secure and compliant with regulatory requirements.
- User training: Detailing the training program for personnel who will interact with the AI system ensures training aligns with regulatory standards and that operators are proficient in using the AI tool effectively and safely.
- Lifecycle description: Providing a comprehensive description of the AI application's lifecycle, including data collection, model development and validation processes assist in addressing aspects like continuous training versus locked models, performance monitoring and retraining.
- Continuously training versus locked model: It is good practice to explain whether the AI model is continuously updated or if it remains static (locked) and to describe the rationale and any safety implications.
- Performance monitoring: Outlining how performance is monitored, including metrics and thresholds and indicating how deviations from acceptable performance are addressed and if performance degradation becomes unacceptable, retraining is essential. However, this retraining process should not affect the licensing basis of the tool unless there are significant changes to its functionality or its safety impact.
- Data:
  - Dynamic model versus fixed model: Distinguishing between dynamic models that operate on real-time collected data and fixed models that use static datasets assist in explaining how the choice impacts system behaviour and safety. As well, specifying the impact of the loss of a data stream and describing how the AI application reacts to data stream interruptions or failures to ensure safe operation builds helps achieve a higher level of confidence in the overall system.
  - Characterization of data during lifecycle: When detailing how data used for AI model training and operation is characterized during its lifecycle it is important to address aging or failing sensors, data quality and any data pre-processing or cleansing.
  - Data sharing with regulator versus characterization of data: Defining whether data is shared with the regulator or if data characterization and relevant information are provided instead helps identify how the regulator can access data if necessary for assessments and clearly communicates to the regulator how the AI tool is applicable, its limitations and how it aligns with regulatory standards and requirements.

In conclusion, engaging with the regulator with a well-prepared and transparent approach is essential to ensure the safe and compliant use of AI in nuclear power plants. It involves early communication, robust implementation and a clear understanding of data and safety implications.

5.9.SAFETY AND SECURITY IMPLICATIONS

IAEA safety standard “Safety of Nuclear Power Plant design” (SSR-2/1, Rev1) requirement 62 states “Instrumentation and control systems for items important to safety at the nuclear power plant shall be designed for high functional reliability and periodic testability commensurate with the safety function(s) to be performed.” And requirement 63 states “If a system important to safety at the nuclear power plant is dependent upon computer-based equipment, appropriate standards and practices for the development and testing of computer hardware and software shall be established and implemented throughout the service life of the system, and, in particular, throughout the software development cycle. The entire development shall be subject to a quality management system.”

Due to the complex algorithms used and the extensive amount of data involved in the analyses, the reliability of AI applications cannot be evaluated through traditional software verification and validation methods alone. One of the most crucial steps in traditional software verification and validation is ensuring that functional requirements are exactly implemented in the software. However, for AI applications, functionality is not implemented solely based on requirements. Therefore, the conventional verification methods may not guarantee the software's reliability.

Also, AI applications can provide advanced functions that were not previously used in NPPs. Even for the AI application in non-safety related system, it should be demonstrated that the failure of an AI-based application would not interfere with safety functions.

6. SUMMARY AND CONCLUSION

This publication presented benefits of AI along with high-level considerations and guidance that could enable AI deployment across several applications in the nuclear power industry. In this section, few key take aways or lessons learned are summarized.

There is currently quite high interest in AI/ML and the new opportunities offered by this technology. There are myriad reasons to consider deploying AI as a tool to assist humans operating nuclear plants. They range from enhancing safety and security, reducing maintenance costs, potentially aiding operators in accident conditions as well as other beneficial reasons. At the same time, the wide general media coverage around these technologies may build expectations or concerns that are not pertinent to the specific application being considered. Experience and example demonstration discussed in this publication has shown that it is important to facilitate sufficient communication among the stakeholders to explicate the expected benefits and limitations of the AI/ML application and to focus on possible concerns relevant to the application.

The role of subject matter expertise and domain knowledge in conjunction with the AI/ML expertise is emphasized as a key success factor on several applications. Experience with engineering systems is that AI methods are made more robust and reliable if domain knowledge is embedded in the algorithm, in this case physics information. Many engineering problems in the light water reactors are high value but are resistant to solution by traditional engineering methods, e.g., [3]. ML coupled with subject matter expertise can provide a solution to these otherwise intractable problems. Some of the examples discussed in this publication have highlighted that the role of subject matter experts, have proven decisive in developing successful models with high reliability. The role is important both in preparation of high quality training data during model development and then in evaluation of the model performance in the application context.

The development of the NDE data analysis has shown the importance of building the models and applications incrementally and providing the users opportunity to build trust in the models. Throughout the development, EPRI has reported the results in industry events and shown applicability in field trials. Utility personnel, inspection vendors and regulator representatives have had the opportunity to witness and monitor the development and raise potential issues to be resolved prior to adoption. As the models near field application, the role of industry guidance increases. Especially with a new technology, like AI/ML, the value of best practices and industry level guidance, such as EPRI MRP documents [65] and European network for inspection and qualification (ENIQ) recommended practices [66] help individual licensees adopt the new technologies in a safe and controlled way.

In this publication, it was discussed that AI/ML applications would benefit from a lifecycle approach as they are likely to involve multiple stakeholders’ engagements, require interdisciplinary teams to develop a solution, and require special infrastructure considerations (integrating with legacy plant systems) that might require long-term monitoring and maintenance. The degree to which an AI lifecycle should be formalized and adhered to depends on the safety, operational, and business risks posed by the application. Parties interested in the deployment of AI should perform a thorough technical, economic, and risk assessment before initiating AI development and deployment initiatives. As with other technology deployment efforts in the nuclear power industry, it is expected that a deliberate, planned, and methodical deployment of AI applications will yield the highest likelihood of success.

Several risks and challenges associated with AI are present as in other types of mature technologies. As such, wherever possible, existing standards and guidance should be adapted and applied to AI application deployment.

One of the important aspects of AI/ML development and deployment is the need for heterogeneous and relevant data, as AI techniques are almost always data centric. As noted, the success of AI applications relies largely on the characteristics and proper use of the underlying data used in development. It is recognized that, typically, large data amounts are needed and for the development of AI applications, where it is generally separated in three distinct groups (training, validation and testing) based on how it is used. While a variety of data categories exist (and often multiple need to be leveraged to meet the necessary data volume), some categories are more or less appropriate for use as part of each of those groups, so it is important to understand the data source and its usage during development.

Given the central importance of development data towards the ultimate performance of AI applications, the quality of that data is a key. Indeed, often the larger portion of development time is used in data curation. Several different aspects of data are discussed that, together, make up the concept of data quality. In particular, it is paramount to have good alignment between the characteristics of the data used in model development and those of the data in the target application scenario, since the overall characteristics of the data used in development (and specially during the testing phase) define and bind the model's range of applicability: AI applications do not, typically, lend themselves to assumptions of generalization or extrapolations.

Beyond the initial development, data management practices and overall data quality management are often needed throughout the application's life. While some of these are similar to any other applications, such as having proper cyber and physical security practices as well as adequate storage solutions in place, specific actions may be needed to address vulnerabilities that are specific or more significant to some AI applications, such as data drift and data poisoning.

Lastly, it is recognized that in many cases the large volume of data that properly encompasses the desired scope of application in the industry is unlikely to be obtained by any single organization. In such cases, data sharing amongst the industry may be the only feasible approach to leverage some of the benefits AI solutions can bring to the industry. While there are valid concerns about data sharing, there are ways to address or minimize many of them that could enable the realization of the value and benefit that data sharing can bring to the industry.

There are myriad reasons to consider deploying AI as a tool to assist humans operating nuclear plants. They range from enhancing safety and security, reducing maintenance costs, potentially aiding operators in accident conditions as well as other beneficial reasons.

As of the publication of this paper, there is little in place regarding the regulation of AI used at nuclear plants. However, nuclear regulators are making strides to understand AI and the implications its deployment may have on the regulation of nuclear plants and two examples discussed in this chapter highlight the direction nuclear regulators are headed. As with all activities at nuclear plants, it is expected regulators will apply a graded approach to the deployment that the regulator whereby the level of oversight is commiserated with the potential risk of implementing the given AI into the facilities operation.

For operators of nuclear plants, it is important that prior to considering implementation of AI, for successful deployment and future operation, that all elements of the AI and its interfaces with the plant and humans charged with its safe operation are characterized and considered. Staff need to be trained, hardware and software must be in place as needed to allow the AI to fulfil its intended function. It is expected that the level of autonomy and the application, particularly in safety critical applications, will drive discussion with the regulator and dictate the scrutiny brought to the specific AI implementation.

Early engagement with all stakeholders, including the regulator, and the establishment of multi-disciplinary teams at the onset of development is good practice. Explainability of the AI and ongoing validation of its function and results will be required to assist all stakeholders in accepting and understanding the AI, including any limitations it may have.

In summary, this publication has discussed different and important aspects of deployment of AI/ML technologies in the nuclear power industry. Significant progress has been made in recent past years, however, there are several challenges and opportunities before we see successful and safe usage of AI/ML in nuclear power industry.

REFERENCES

[1] INTERNATIONAL ORGANIZATION FOR STANDARDIZATION, ISO/IEC, Artificial Intelligence-Data Quality for Analytics and Machine Learning Part 1: Overview, terminology and examples, ISO/IEC DIS 5259-1, Geneva (2023).

[2] INTERNATIONAL ORGANIZATION FOR STANDARDIZATION, ISO/IEC, Information Technology - Artificial Intelligence - Data Life Cycle Framework, ISO/IEC 8183, Geneva (2023).

[3] NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY, Recommendation for Key Management, NIST Special Publication 800-57 Part 1 – General, Gaithersburg (2020).

[4] MAJUMDAR, M.C., MAJUMDAR, D., SACKETT, J.I., Artificial Intelligence and Other Innovative Computer Applications in the Nuclear Industry (October 1988).

[5] US Department of Energy, AI Risk Management Playbook (AI RMP) https://www.energy.gov/ai/doe-ai-risk-management-playbook-airmp (2022).

[6] LUTKEVICH, B., “What is generative design?”, https://www.techtarget.com/whatis/definition/generative-design (2023, July 10).

[7] DENNIS, H., “NASA's "evolved structures" radically reduce weight – and waiting”, New Atlas, https://newatlas.com/space/nasa-generative-design/ (2023).

[8] KIM, S.G., SEONG, P.H., CHAE, Y. H., “Development of a generative-adversarial-network-based signal reconstruction method for nuclear power plants”, Annals of Nuclear Energy, Volume 142, July 2020, 107410 (2020). https://www.sciencedirect.com/science/article/pii/S0306454920301080,

[9] MARTINEAU, K., “Want cheaper nuclear energy? Turn the design process into a game”, MIT News, Massachusetts Institute of Technology (2020). https://news.mit.edu/2020/want-cheaper-nuclear-energy-turn-design-process-game-1217

[10] NGUYEN, T.N., VILIM, R., "Direct Bayesian Inference for Quantitative Model-Based Fault Detection and Diagnosis", Annals of Nuclear Energy, Volume 194, December 2023, 109932 (2023).

[11] VILIM, R., GRELLE, A., BORING, R., THOMAS, K., ULRICH, T., LEW, R.,” Computerized Operator Support System and Human Performance in the Control Room”, 10th International Topical Meeting on Nuclear Plant Instrumentation, Control and Human Machine Interface Technologies, San Francisco, CA, June 11-15, 2017.

[12] WANG H., GRUENWALD, J., TUSAR, J., VILIM, R. “Moisture-carryover performance optimization using physics-constrained machine learning”, Progress in Nuclear Energy, 135, May 1, 2021.

[13] MOISEYTSEVA, V., GRABASKAS, D., PONCIROLI, R., NGUYEN, T., “Cost-Benefit Analyses through Integrated Online Monitoring and Diagnostics,” ANL/NSE-23/22, Argonne National laboratory, March 2023.

[14] PONCIROLI, R., MOISEYTSEVA, V., DAVE, A.J., NGUYEN, T.N., VILIM, R.B., “Design and Prototyping of Advanced Control Systems for Advanced Reactors Operating in the Future Electric Grid,” Argonne National Laboratory, ANL/NSE-23/48, August 2023.

[15] BORELLA A., ROSSA, R., ZAIOUN, H., “Determination of 239Pu content in spent fuel with the SINRD technique by using artificial and natural neural networks”, ISSN: 1977-5296, DOI: 10.3011/ESARDA.IJNSNP.2019.5, ESARDA Bulletin, Volume 58, June 2019.

[16] ROSSA R., BORELLA A., GIANI, N., “Comparison of machine learning models for the detection of partial defects in spent nuclear fuel”, Annals of Nuclear Energy, Volume 147, November 2020, 107680 (2020).

[17] AL-DBISSI, M., ROSSA, R., BORELLA, A., PÁZSIT, I., VINAI, P. “Identification of diversions in spent PWR fuel assemblies by PDET signatures using Artificial Neural Networks (ANNs)”, Annals of Nuclear Energy, Volume 193, December 2023, 110005 (2023).

[18] GRAPE, S., BRANGER, E., ELTER, Z., PÖDER BALKESTÅHL, L., “Determination of spent nuclear fuel parameters using modelled signatures from non-destructive assay and Random Forest regression”, Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, Volume 969, July 2020.

[19] HELLESEN, C., GRAPE, S., JANSSON, P., JACOBSSON SVÄRD, S., ÅBERG LINDELL, M., ERSSON, P., "Nuclear spent fuel parameter determination using multivariate analysis of fission product gamma spectra", Annals of Nuclear Energy, Volume 110, pages 886–895, December 2017.

[20] BACHMANN, A. M., COBLE, J. B., SKUTNIK, S. E., "Comparison and uncertainty of multivariate modeling techniques to characterize used nuclear fuel", Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, Volume 991, March 2021.

[21] BORELLA, A., ROSSA, R., TURCANU, C., "Signatures from the spent fuel: simulations and interpretation of the data with neural network analysis", ISSN: 1977-5296, DOI: 10.3011/ESARDA.IJNSNP.2017.15, ESARDA Bulletin, Volume 49, no. 31, December 2017.

[22] MISHRA, V., BRANGER, E., ELTER, Z., GRAPE, S., JANSSON, P., "Comparison of different supervised machine learning algorithms to predict PWR spent fuel parameters", Institute of Nuclear Materials Management (2021).

[23] SHOMAN, N., CIPITI, B., "Advances in machine learning for safeguarding a PUREX reprocessing facility", SAND2020-5394C, Sandia National Laboratories, New Mexico, 2020

[24] NUCLEAR REGULATORY COMMISSION, U.S. Code of Federal Regulations (CFR), "Changes, tests, and experiments", Section 59, Part 50, Chapter 1, Title 10, "Energy", 10 CFR Part 50.59.

[25] LANGNER R., "Stuxnet: Dissecting a cyberwarfare weapon", doi: 10.1109/MSP.2011.67, *IEEE Security and Privacy*, vol. 9, no. 3, pp. 49–51, 2011.

[26] Japan Today, "Monju power plant facility PC infected with virus," https://japantoday.com/category/national/monju-power-plant-facility-pc-infected-with-virus (current as of Jan. 14, 2022).

[27] STEITZ C., AUCHARD E., "German nuclear plant infected with computer viruses, operator says," https://www.reuters.com/article/us-nuclearpower-cyber-germany/german-nuclear-plant-infected-with-computer-viruses-operator-says-idUSKCN0XN2OS (current as of Jan. 14, 2022).

[28] IVANOV E., DE SAINT-JEAN C., SOBES V., "Nuclear data assimilation, scientific basis and current status", EPJ Nuclear Science and Technologies, Volume 7, no. 9, 2021.

[29] DOMINESEY K. A., JI W., "Reduced-order modeling of neutron transport separated in space and angle via proper generalized decomposition", Journal of Computer Physics, Volume 449, January 2022.

[30] LI Y., BERTINO E., ABDEL-KHALIK H., "Effectiveness of Model-Based Defenses for Digitally Controlled Industrial Systems: Nuclear Reactor Case Study", Nuclear Technology, Volume 206, no. 1, pp. 82–93, 2018.

[31] ZHANG F., COBLE J. B., "Robust localized cyber-attack detection for key equipment in nuclear power plants," Progress in Nuclear Energy, Volume 128, 2020.

[32] SUNDARAM A., ABDEL-KHALIK H. S., ASHY O., "A data analytical approach for assessing the efficacy of Operational Technology active defenses against insider threats", Progress in Nuclear Energy, Volume 124, 2020.

[33] MO Y., SINOPOLI B., "Secure control against replay attacks", 47th Annual Allerton Conference on Communication, Control, and Computing, pp. 911–918, doi: 10.1109/ALLERTON.2009.5394956, 2009.

[34] FLETCHER J., "Deepfakes, Artificial Intelligence, and Some Kind of Dystopia: The New Faces of Online Post-Fact Performance," Theatre Journal, Volume 70, no. 4, pp. 455–471, doi: 10.1353/tj.2018.0097, 2018.

[35] TANCIK M., MILDENHALL B., and NG R., "StegaStamp: Invisible Hyperlinks in Physical Photographs", Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2114–2123, doi: 10.1109/CVPR42600.2020.00219, April 2019.

[36] WANG R., JUEFEI-XU F., LUO M., LIU Y., WANG L., “FakeTagger: Robust Safeguards against DeepFake Dissemination via Provenance Tracking”, MM 2021 - Proceedings of the 29th ACM International Conference on Multimedia, pp. 3546–3555, doi: 10.1145/3474085.3475518, September 2020.

[37] SUNDARAM A., ABDEL-KHALIK H., “Covert Cognizance: A Novel Predictive Modeling Paradigm,” Nuclear Technology, Volume 207, no. 8, pp. 1163-1181, doi: 10.1080/00295450.2020.1812349, 2021.

[38] GOODFELLOW I. J., SHLENS J., SZEGEDY, C., “Explaining and harnessing adversarial examples,” arXiv, 2015.

[39] EYKHOLT K. et al., “Robust Physical-World Attacks on Deep Learning Visual Classification,” arXiv, April 2018.

[40] JONES E. et al., “Robust Encodings: A Framework for Combating Adversarial Typos”, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 2752–2765,doi: 10.18653/v1/2020.acl-main.245, May 2020.

[41] GOPALAKRISHNAN S. et al., “Combating Adversarial Attacks Using Sparse Representations”, 6th International Conference on Learning Representations - Workshop Track Proceedings, Mar. 2018.

[42] SHOKRI R., STRONATI M, SONG C, SHMATIKOV V., “Membership Inference Attacks Against Machine Learning Models”, arXiv, March 2017.

[43] SUNDARAM A., ABDEL-KHALIK H. S., AL RASHDAN A., “Deceptive Infusion of Data (DIOD) for Nuclear Reactors”, Transactions of the American Nuclear Society, Volume 125, no. 1, pp.264-266, 2021.

[44] BRUNDAGE M. et al., “The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation,” arXiv, February 2018.

[45] OpenAI et al., “Dota 2 with Large Scale Deep Reinforcement Learning”, arXiv, Dec. 2019.

[46] RISI S., PREUSS M., “Behind DeepMind’s AlphaStar AI that Reached Grandmaster Level in StarCraft II”, KI - Künstliche Intelligenz, Volume 34, no. 1, pp. 85–86, doi: 10.1007/S13218-020-00642-1, February 2020.

[47] CHEN J. X., “The Evolution of Computing: AlphaGo”, Computing in Science and Engineering, Volume 18, no. 4, pp. 4–7, doi: 10.1109/MCSE.2016.74, July 2016.

[48] MCCARTHY J., “Some expert systems need common sense”, Annals of the New York Academy Sciences, Volume 426, 1984.

[49] BROWN T. B. et al., “Language Models are Few-Shot Learners”, Proceedings of Advances in Neural Information Processing Systems 33, 2020.

[50] GRIET, M., VENTURINI, V., FLURY, C., Sistema Avanzado de Alarmas para el Reactor RA6C, Centro Atómico Bariloche. Comisión Nacional de Energía Atómica. Argentina, 2016. https://inis.iaea.org/collection/NCLCollectionStore/_Public/51/100/51100291.pdf

[51] Alarm systems - a guide to design, management and procurement, Engineering Equipment and Materials Users Association, EEMUA Publication 191, Second Edition.

[52] SANCHEZ-PI, N., LEME, L.A.P., GARCIA, A.C.B., Intelligent Agents for Alarm Management in Petroleum Ambient, p 43 – 53, 1 January 2015.

[53] GAGGIOLI, A., User’s experience in ambient intelligence systems. Ambient Intelligence: Conceptual and Practical Issues, p 100–110, 2005.

[54] VENTURINI, V., Sistema multi-agente basado en contexto, localización y reputación para dominios de inteligencia artificial, PhD thesis. Carlos III University of Madrid., Spain, 2012.

[55] FOSTER-ROMAN, D., LEUNG, J., METZLER, J., NAQVI, R., SHAHBANDI, N., ZARZECZNY, W., "Ontario Power Generation's Monitoring & Diagnostic Centre", Proc. 39th Annual Conf. of the Canadian Nuclear Society and 43rd Annual CNS/CNA Student Conf., Ottawa, ON, Canada, IAEA International Nuclear Information System, (2019).

[56] HRYNIEWICKI, M.K., STRAPP, J., "Improving Nuclear Unit Outage Scheduling with Artificial Intelligence", Power Engineering, https://www.power-eng.com/nuclear/improving-nuclear-unit-outage-scheduling-with-artificial-intelligence (2019).

[57] INSTITUTE OF NUCLEAR POWER OPERATORS, Performance Continuum Manual, Rev. 3.1, (2023).

[58] ELECTRIC POWER RESEARCH INSTITUTE, AI-Assisted Analysis of Ultrasonic Inspections, Rep. 3002023718EPRI, Palo Alto, 2022.

[59] ELECTRIC POWER RESEARCH INSTITUTE, AI Tool Developed by EPRI Significantly Cuts Analysis Time in U.S. Nuclear Plant Field Trial: Data Analysis Takes Four Hours Compared to Four Days Without Artificial Intelligence, Rep. 3002025510, EPRI, Palo Alto, 2022.

[60] ELECTRIC POWER RESEARCH INSTITUTE, Quick Insight Brief: Using Artificial Intelligence to Maximize the Benefits of Drones for Nuclear Power Plants, Rep. 3002023930, EPRI, Palo Alto, 2022.

[61] ELECTRIC POWER RESEARCH INSTITUTE, Automating Corrective Action Programs in the Nuclear Industry, Rep. 3002023821, EPRI, Palo Alto, 2022.

[62] ELECTRIC POWER RESEARCH INSTITUTE, Quick Insight Brief: Leveraging Artificial Intelligence for the Nuclear Energy Sector, Rep. 3002021067, EPRI, Palo Alto, 2021.

[63] ELECTRIC POWER RESEARCH INSTITUTE, Quick Insight Brief: Leveraging Artificial Intelligence for Nondestructive Evaluation, Rep. 3002021074, EPRI, Palo Alto, 2021.

[64] ELECTRIC POWER RESEARCH INSTITUTE, Automated Analysis of Remote Visual Inspection of Containment Buildings, Rep. 3002018419, EPRI, Palo Alto, 2020.

[65] ELECTRIC POWER RESEARCH INSTITUTE, Materials Reliability Program: Guideline for Nondestructive Examination of Reactor Vessel Upper Head Penetrations, Revision 1 (MRP-384), Rep. 3002017288, EPRI, Palo Alto, 2019.

[66] VIRKKUNEN, I., BOLANDER, M., MIORELLI, R., JOHANSSON, O., KICHERER, P., CURTIS, C., MARTIN, O., ENIQ Recommended Practice 13: Qualification of Non-Destructive Testing Systems that Make Use of Machine Learning. European Network for Inspection & Qualification, ENIQ Report No. 65 (2021).

[67] PARK, N., MOHAMMADI, M., GORDE, K., JAJODIA, S., PARK, H., KIM, Y., Data synthesis based on generative adversarial networks. arXiv preprint arXiv:1806.03384 (2018).

[68] TYYSTJÄRVI, T., VIRKKUNEN, I., FRIDOLF, P., ROSELL, A., BARSOUM, Z., Automated defect detection in digital radiography of aerospace welds using deep learning. Springer Science and Business Media LLC. Welding in the World, 66. Welding in the World, 66(4), 643-671. doi:10.1007/s40194-022-01257-w (2022).

[69] VIRKKUNEN, I., KOSKINEN, T., JESSEN-JUHLER, O., Virtual round robin–A new opportunity to study NDT reliability. Elsevier. Nuclear Engineering and Design, 380. Nuclear Engineering and Design, 380, 111297, (2021).

[70] SILJAMA, O., KOSKINEN, T., JESSEN-JUHLER, O., VIRKKUNEN, I., Automated Flaw Detection in Multi-channel Phased Array Ultrasonic Data Using Machine Learning. Springer Science and Business Media LLC. Journal of Nondestructive Evaluation, 40. Journal of Nondestructive Evaluation, 40(3). doi:10.1007/s10921-021-00796-4 (2021).

[71] VIRKKUNEN, I., KOSKINEN, T., JESSEN-JUHLER, O., RINTA-AHO, J., Augmented Ultrasonic Data for Machine Learning. Springer Science and Business Media LLC. Journal of Nondestructive Evaluation, 40. Journal of Nondestructive Evaluation, 40(1). doi:10.1007/s10921-020-00739-5 (2021).

[72] SEUACIUC-OSORIO, T., ESP, L., AI-Assisted Analysis of Ultrasonic Inspections. Technical brief, 3002023718 (2022).

[73] INTERNATIONAL ATOMIC ENERGY AGENCY, Computer Security Techniques for Nuclear Facilities, IAEA Nuclear Security Series No. 17-T (Rev. 1), IAEA, Vienna (2021).

[74] INTERNATIONAL ATOMIC ENERGY AGENCY, Security of Nuclear Information, IAEA Nuclear Security Series No. 23-G, IAEA, Vienna (2015).

[75] MITRE ATLAS, Available online at https://atlas.mitre.org/ Accessed December 2023.

[76] METZEN, J. H., KUMAR, M. C., BROX T., FISCHER, V., "Universal Adversarial Perturbations Against Semantic Image Segmentation," 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2774-2783, doi: 10.1109/ICCV.2017.300, Venice, Italy (2017).

[77] HU, W., TAN, Y., Generating Adversarial Malware Examples for Black-Box Attacks Based on GAN. ArXiv, abs/1702.05983 (2017).

[78] BAI, T., LUO, J., ZHAO, J., WEN, B., WANG, Q., “Recent Advances in Adversarial Training for Adversarial Robustness.” International Joint Conference on Artificial Intelligence, IJCAI-21 (2021).

[79] YANG, Q., LIU, Y., CHEN, T., TONG, Y., Federated Machine Learning: Concept and Applications, ACM Transactions on Intelligent Systems and Technology, Volume 10, Issue 2, Article 12, https://doi.org/10.1145/3298981 (March 2019).

[80] INTERNATIONAL ELECTROTECHNICAL COMMISSION, Nuclear power plants – Instrumentation, control and electrical power systems – Security controls, IEC 63096:2020, IEC, Geneva (2020).

[81] INTERNATIONAL ATOMIC ENERGY AGENCY, Computer Security Approaches to Reduce Cyber Risks in the Nuclear Supply Chain, IAEA-TDL-011, IAEA Non-serial Publications, IAEA, Vienna (2022).

[82] LEE J., Human factors and ergonomics in automation design, in Handbook of Human Factors and Ergonomics Ed. by Salvendy G., John Wiley & Sons (2006).

[83] ENDSLEY M., KIRIS E., The out-of-the-loop performance problem and level of control in automation, Human Factors, 37(2):381–394 (1995).

[84] SHEIDERMAN B., Human-Center AI, Oxford University Press (2021).

[85] RIEDL M., Human-centered artificial intelligence and machine learning, Human Behavior and Emerging Technologies, 1(1):33-36 (2019).

[86] XU W., Toward human-centered AI: A perspective from human-computer interaction, Interactions, 26(4):42-46 (2019).

[87] DAFOE A. ET AL., Open Problems in Cooperative AI, https://doi.org/10.48550/arXiv.2012.08630 (2020).

[88] DAFOE A. ET AL., Cooperative AI: machines must learn to find common ground, Nature, 593:33-36 (2021).

[89] AGBAJI, D., LUND, B., & MANNURU, N. R., Perceptions of the Fourth Industrial Revolution and Artificial Intelligence Impact on Society. arXiv preprint arXiv:2308.02030. Retrieved from https://arxiv.org/abs/2308.02030 (2023).

[90] KOTTER, J.P., Leading Change. Boston, MA: Harvard Business School Press (1996).

[91] Lewin, K., Group Decision and Social Change. In: Maccoby, E.E., Newcomb, T.M. and Hartley, E.L., Eds., Readings in Social Psychology, Holt, Rinehart, Winston, New York, 197-211 (1958).

[92] HIATT, J., ADKAR: A Model for Change in Business, Government and our Community. Loveland, CO: Prosci Learning Center Publications (2006).

[93] DALE CARNEGIE & ASSOCIATES, Preparing People for Success with Generative AI. Retrieved from https://www.dalecarnegie.com/en/resources (2023).

ANNEX I.

FRAMEWORK OF INTELLIGENT NUCLEAR POWER PLANTS

For long-term development of intelligent nuclear power plants, it is necessary to coordinate the following aspects:

- *Intelligent applications:* Choosing appropriate intelligent algorithms (sometimes combining traditional analysis and intelligent computing) to develop AI technology applications that match the functional requirements of NPPs users.
- *Data:* Fully mining the existing data of NPPs and importing additional effective data to ensure the development and validation of AI technology applications.
- *Platform:* Overall planning of platform construction, with high scalability, accommodating various intelligent applications, effectively promoting high data integration and reducing project deployment costs.

I-1. FUNCTIONAL OBJECTIVES OF INTELLIGENT NUCLEAR POWER PLANTS

In the near to medium term, the application of AI technology can contribute to the following functional objectives:

- *Intelligent operation:* Reduce the workload of operator monitoring and avoid human errors through real-time monitoring and early warning of plant operating status; Reduce the workload of on-site O&M personnel and improve efficiency through intelligent monitoring and analysis of equipment performance; Optimize the execution cycle of regular tests by accurately understanding the equipment status.
- *Intelligent equipment management and maintenance:* By utilizing technologies such as equipment status monitoring, fault diagnosis and remaining useful life prediction, the operational status of structures/systems/components (SSCs) can be fully grasped, condition-based maintenance decisions can be made, early warning will be given before equipment failures and SSCs can be utilized more economically.
- *Intelligent hazard prevention:* Improve the detection and identification capabilities of hazardous factors in and around NPPs and rely less on manual inspections.
- *Intelligent emergency response:* Nuclear accidents have wide range and long duration of impacts and it is necessary to provide nuclear emergency decision-making support based on massive information in a short period of time.

I-2. INTELLIGENT NPP APPLICATIONS ARCHITECTURE

To further use the operational data of NPPs, a reliable and flexible system architecture should be created to provide the relevant data for intelligent applications. The architecture should not affect the normal operation of the plant and the data should be used in a secure and efficient way.

As is shown in FIG. I-1, **the**e operator controls the plant in the main control room and the distributed control system (DCS) executes the command; The intelligent operation assistance system receives and analyses the plant data from DCS; Data analytics are presented to the operator through the intelligent monitoring centre and assist the operator in making operational decisions; The intelligent emergency system receives plant data in one direction to execute emergency related functions and data analytics are used by the intelligent emergency centre to make decisions; The intelligent operation management system deployed in management information area receives plant data from production control area in one direction. This system architecture and data transmission method can reduce the investment in platform construction and later operation and maintenance in order to comprehensively utilize power plant data and apply AI technologies.

I-3. PLATFORM OF AI APPLICATIONS

Build an integrated information platform for AI applications in NPPs. The platform is an important core infrastructure for nuclear power. It can provide power plants with unified data collection or access services, business application resources (computing power, storage, memory) services, network security protection services and other functions. The platform can realize flat application of data at unit level. The platform can efficiently deploy various intelligent application functions and comprehensively improve the intelligent unit operation. The platform can further process the unit-level operation and management data of the unit level and send it to the plant-level management system, further enhancing the application value of the core data for multi-unit and multi-plant management.

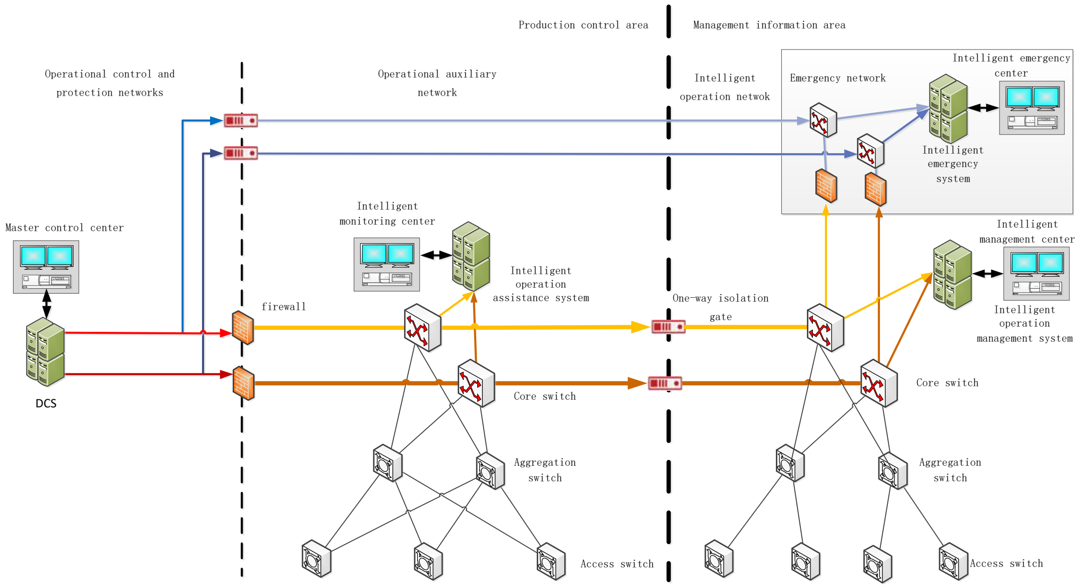

Production control area
Management information area
Operational control and protection networks
Operational auxiliary network
Intelligent operation netwok
Emergency network
Intelligent emergency center
Intelligent emergency system
Master control center
Intelligent monitoring center
Intelligent operation assistance system
Intelligent management center
Intelligent operation management system
firewall
One-way isolation gate
DCS
Core switch
Core switch
Aggregation switch
Aggregation switch
Access switch
Access switch

FIG. I-1. Example of possible architecture of intelligent nuclear power plant

GLOSSARY

Autonomy.

convolutional neural networks.

data corruption. Refers to errors unintentionally introduced during writing, reading, storage, transmission or processing of data that may render the data unreadable or otherwise compromised.

data bias. Systematic errors or skewed distributions in the data that make it favour one class or category over another.

data distribution. Refers to the way data points are spread across the relevant ranges, categories, or classes.

data drift. Unexpected, undocumented or uncontrolled changes over time in the data characteristics, such as data distribution.

data governance. The overall practices and processes for suitable data management, including provisions for how data is properly recorded, processed, retained, used and distributed throughout its lifecycle.

data integrity. Refers to the property that data maintains its quality throughout its lifecycle, not deteriorating by unauthorized agents or processes, intentionally or unintentionally.

data leakage. Refers to a failure in proper data partitioning where data or information about data used in training and validation are intentionally or unintentionally included in test sets.

data poisoning. The intentional injection of malicious data into AI/ML models either during training and development or inference at deployment.

data quality. Refers to the overall condition of the combined set of relevant data characteristics data make a dataset suitable for a given application.

data scope. Refers to the context of and range covered by a dataset.

data security. Refers to the overall practices and effects of ensuring the maintenance of proper data integrity, confidentiality and availability.

data sharing. Refers to the practice of data owners sharing data with and making it available to a broader community.

deep learning.

dynamic models. Applications that include a feedback loop where the deployed model continuously learns from current data and changes as adequate. In this case, the current data is actively and continuously being included in the training of the model, which is periodically and automatically updated based on the new information.

explainability.

field data. Data collected from sensors or humans within the operation environment of a nuclear facility.

guardrails.

laboratory data. Data gathered in controlled settings to conduct experiments and physical simulations.

natural language processing.

open-source data. Publicly available data.

repeatability.

reproducibility.

simulated data. Data generated through numerical or analytical models that rely on an underlying physical description and understanding of the phenomena.

synthetic data. Data generated through methods that do not rely on an underlying physical model describing the physical phenomena, such as data generated by artificial intelligence methods such as generative adversarial networks.

virtual data. Data generated by designed manipulation of existing real field or laboratory data.

ABBREVIATIONS

AI	artificial intelligence
ANN	artificial neural network
AmI	environmental intelligence
DAS	??
DT	digital twin
FOL	first-order logic
GAN	generative adversarial network
H&S	health and safety
HCAI	human-centered ai
IoT	internet of things
LLM	large language model
M&D	monitoring and diagnostics
MAS	multi-agents systems
ML	machine learning
NDA	non-destructive assay
NDE	non-destructive evaluation
NLP	natural language processing
O&M	operations and maintenance
OEM	original equipment manufacturer
OPEX	operational experience
QA	quality assurance
R&D	research and development
SINRD	self-indication neutron resonance densitometry
SMR	small modular reactors
SNF	spent nuclear fuel
SQA	software quality assurance
T&L	teaching and learning
VVUQ	verification, validation and uncertainty quantification

CONTRIBUTORS TO DRAFTING AND REVIEW

Abdel-Khalik, H.	Purdue University, United states of America
Agarwal, V.	Idaho National Laboratory, United States of America
Al-dbissi, M.	Chalmers University of Technology
Al-Rashdan, A.Y.	Idaho National Laboratory, United States of America
Andrachek, J.	Pressurized Water Reactor Owners Group, United States of America
Ayalasomayajula, S.	Nuclear Promise X, Canada
Betancourt, L.	Nuclear Regulatory Commission, United States of America
Boring, R.	Idaho National Laboratory, United States of America
Briquez, B.	Tecnatom/Westinghouse, Spain
Busquim, R.	International Atomic Energy Agency
Cancila, D.	Commissariat à l'énergie atomique et aux énergies alternatives, France
Carter, C.E.	Utilities Service Alliance, United States of America
Chu, J.	China Nuclear Power Engineering, China
Comeaux. K.	Institute of Nuclear Power Operations, United States of America
Cox, N.	Nuscale, United States of America
Deng, S.	China Nuclear Power Engineering, China
Dennis, M.	Nuclear Regulatory Commission, United States of America
El Bouzidi, S.	Canadian Nuclear Laboratories Ltd., Canada
Foster-Roman, D.	Ontario Power Generation, Canada
Gadallah, I. M.	Egyptian Atomic Energy Authority, Egypt
Gruenwald, J.T.	Blue Wave AI Labs, United States of America
Golightly, C.	Energy Northwest, United States of America
Guerra-O'Hanlon, S.	Adelard - NCC Group, United Kingdom
Hall, M.	Evergy, United States of America
Hathaway, A.	Nuclear Regulatory Commission, United States of America
Hewes, M.	International Atomic Energy Agency
Highland, B.D.	Energy Northwest, United States of America
Kim, J.	Chosun University, Republic of Korea
Lagarde, J.	Metroscope, France
Lee, K.	Canadian Nuclear Safety Commission, Canada
Li, J.	Tsinghua University, China
Li, M.	China Nuclear Power Engineering, China
Li, W.	China Nuclear Power Engineering, China
Lynde, J.	Pressurized Water Reactor Owners Group, United States of America
Nangia, B.	Nuclear Promise X, Canada
Nieto, L.A.	Comisión Nacional de Energía Atómica, Argentina
Movassat, M.	Ontario Power Generation, Canada

Pereira de M. Martins, G	Electronuclear, Brazil
Powell, M.	Pressurized Water Reactor Owners Group, United States of America
Seuaciuc-Osorio, T.	Electric Power Research Institute, United States of America
Sladek, J.	Canadian Nuclear Safety Commission, Canada
Smith, P.	Lancaster University, United Kingdom
Venturini, V	Comisión Nacional de Energía Atómica, Argentina
Vilim, R.	Argonne National Laboratory, United States of America
Virkkunen, I.	Aalto University, Finland
Walker, C.M.	Idaho National Laboratory, United States of America
Xu S.	China Nuclear Power Engineering, China
Yao, W.	China Nuclear Power Engineering, China
Yu, D.	China Nuclear Power Engineering, China
Zeng, Z.C.	Canadian Nuclear Safety Commission, Canada
Zhang, L.	China Nuclear Power Engineering, China

Consultants Meetings

Vienna, Austria: 5–7 July 2023; 4–8 December 2023;

Technical Meeting

Rockville, United States of America: 18–21 March 2024