CONSIDERATIONS FOR DEVELOPING ARTIFICIAL INTELLIGENCE SYSTEMS IN NUCLEAR APPLICATIONS

SEPTEMBER 2024

Canadian Nuclear Safety Commission
UK Office for Nuclear Regulation
US Nuclear Regulatory Commission

CONTENTS

Disclaimer

The information contained in this document is provided without any warranties or guarantees. References to any specific standards, commercial products, processes, or services does not constitute or imply endorsement or recommendation on the part of the Canadian Nuclear Safety Commission (CNSC), United Kingdom Office for Nuclear Regulation (UK ONR), or the United States Nuclear Regulatory Commission (US NRC). Furthermore, this document should not be considered to represent either requirements or guidance on the part of the CNSC, UK ONR, US NRC. Any future guidance or rulemaking on artificial intelligence (AI), if needed, will follow each respective agency's typical processes.

Purpose

This document was developed in furtherance of the participating organizations' missions, including their responsibilities to identify and disseminate information on technologies deployed, or likely to be deployed, in the nuclear industry. The considerations for developing AI systems in nuclear applications discussed in this paper are intended to be principles that all participants in the AI lifecycle should consider as part of development and deployment. As such, the audience includes parties such as AI developers, end users, applicants, licensees, regulators, and regulatory partners.

Acknowledgement

This document was developed collaboratively between the Canadian Nuclear Safety Commission (CNSC), United Kingdom Office for Nuclear Regulation (UK ONR), and the United States Nuclear Regulatory Commission (US NRC). The principal authors included Kevin Lee (CNSC), Andrew White (UK ONR), Daniel Finnigan (UK ONR) and Matt Dennis (US NRC). The authors would like to thank the technical, legal, administrative, and management staff from the respective agencies that contributed to the publication of this trilateral document.

Alex Viktorov (CNSC)
Anthony Valiaveedu (US NRC)
Barry Hogan (UK ONR)
Brian Green (US NRC)
Brian Torrie (CNSC)
Carol Nove (US NRC)
Courtney Williams (UK ONR)
Dave McIntyre (US NRC)
Edward Nakaza (CNSC)
Howard Benowitz (US NRC)
Jesse Seymour (US NRC)
John Sladek (CNSC)
Joshua Kaizer (US NRC)
Justin Sigetich (CNSC)
Keith Dewar (CNSC)
Kim Lawson-Jenkins (US NRC)
Luis Betancourt (US NRC)
Laura Lynne Churchill (CNSC)
Luke Carter (UK ONR)
Natasha DeActis (CNSC)
Pierre-Daniel Bourgeau (CNSC)
Tex Steinfeldt (US NRC)
Tison Campbell (US NRC)
Tom Eagleton (UK ONR)
Trey Hathaway (US NRC)
Victor Hall (US NRC)
Victor Martins (US NRC)

1 INTRODUCTION

This document outlines high-level principles for the deployment of artificial intelligence (AI) as noted by the Canadian Nuclear Safety Commission (CNSC), the United Kingdom's Office for Nuclear Regulation (UK ONR), and the United States of America's Nuclear Regulatory Commission (US NRC). This AI principles paper describes important topics that should be considered when deploying AI to ensure continued safe and secure operation of nuclear facilities, and other uses of nuclear materials. The three regulatory bodies contributing to this paper recognize the importance of thoughtfully considering these principles when developing, reviewing, and deploying AI systems in any of the three jurisdictions. The audience for this paper includes AI developers, end users, applicants, licensees, regulators, and regulatory partners. The principles discussed here are intended to remind all parties that while we encourage beneficial uses of AI, we need to clarify and address the challenges arising from these fast-developing technologies.

This document defines AI to be a range of technologies that can learn from data or experiences to perform tasks that would otherwise require human intelligence. An AI system is a system that contains AI components (e.g. neural networks), typically developed using software tools. Data is used to train the AI system to give the desired output, and independent data is used to test whether the system's performance meets expectations. A wide range of approaches and tools for the development of AI is available and largely bundled into domain usage areas such as computer vision, natural language processing or generative large language models. The underlying architecture of these models and associated data could remain static or continuously evolve based on newly acquired data.

AI could support outcomes that are not readily attainable via non-AI techniques. For example, AI could analyse a much larger volume of data than conventional approaches, potentially allowing better management of plant risks and improved efficiency. Similarly, sufficiently developed AI could accomplish tasks that have until recently only been possible using humans, reducing the need to enter hazardous areas, and potentially reducing error. Additionally, the ability to retrain AI to benefit from updated information provides flexibility to rapidly learn from previous experience to improve over time. The unique capabilities of AI also make feasible activities that are not achievable by either conventional technologies or humans, creating opportunities to directly and indirectly improve safety, security, and efficiency.

Even though each country is operating with a different regulatory framework, it is nevertheless important to examine some foundational tenets all parties should consider when AI is used in nuclear applications. This document is split into several sections, listed below, that the Canadian, UK, and US nuclear regulators consider may be important in managing the risks arising from the use of AI. Note that the order does not indicate importance.

- Use of existing safety and security engineering principles
- Human and organisational factors
- AI architecture
- AI lifecycle management
- Documenting AI safety and security

2 COUNTRY-SPECIFIC REGULATORY PHILOSOPHIES AND PERSPECTIVES

The CNSC, UK ONR and US NRC collectively prioritize nuclear safety as our primary goal, but the regulatory activities, frameworks, and philosophies underpinning this goal may differ between regulators. Differences in national contexts and priorities reflect the unique combination of historical, political, societal, economic, technological, and environmental factors that shape each country's approach to nuclear regulation. Prior experience has shown that each regulator will conduct its own regulatory analysis and research when presented with innovative technologies such as advanced reactor designs, small modular reactors (SMRs) and digital instrumentation and controls (I&C). However, we recognise that while each country's regulatory framework and processes differ, there are many benefits in sharing knowledge and experience and seeking common technical evaluation approaches for the deployment of AI in nuclear applications.

If one regulator accepts deployment of AI in a given situation, other regulators will not necessarily accept deployment of a similar use, and each may wish to conduct their own assessment. However, the existing relationship between the CNSC, UK ONR, and US NRC through memorandums of understanding and close collaboration on nuclear safety and security topics indicates that the three regulatory bodies will work to harmonize their regulatory approach to the extent practicable. The three regulators are in close collaboration to address the issues surrounding adoption of AI in the nuclear industry, share lessons learned on regulatory activities related to AI, and foster information sharing which facilitates safe and secure use of AI under their respective regulatory mandates.

A description of each country's regulatory framework can be found at the links provided below:

- Canadian Nuclear Safety Commission
- United Kingdom Office for Nuclear Regulation
- United States Nuclear Regulatory Commission

3 HIGH LEVEL CATEGORIES FOR AI USE CASES IN NUCLEAR APPLICATIONS

For the purposes of this paper, we consider there to be four broad categories of AI systems that can be described across the following two axes, significance of AI failure and amount of AI autonomy.[1] This is presented in Figure 1 as a four-region model.

As shown on the horizontal axis of Figure 1, the amount of AI autonomy reflects the progression from higher levels of human involvement to greater machine independence. These two regions reflect the transition from where human decision-making is assisted or augmented by an AI to one where AI decision-making is supervised by a human or little to no human intervention is expected. Therefore, these two regions are referred to as Insight/Collaboration on the left-hand side of the horizontal axis and Operation/Full Autonomy on the right-hand side of the horizontal axis. Further discussion of these terms can be found in the US NRC's AI Strategic Plan (NUREG-2261).

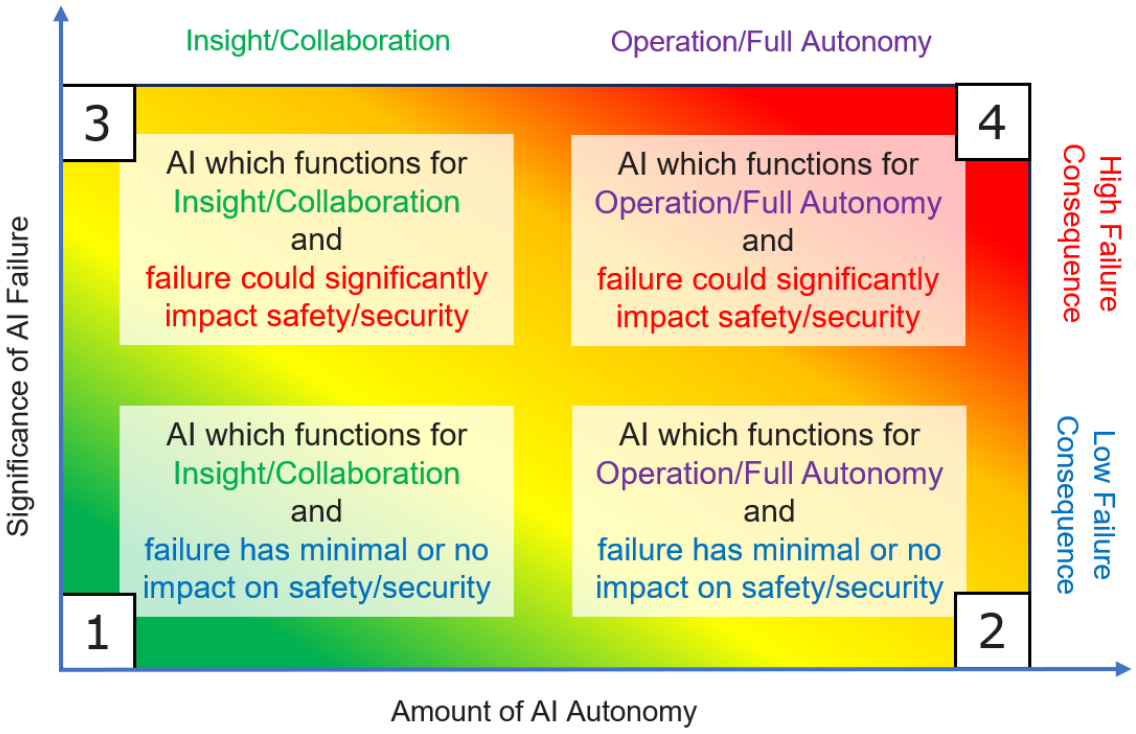


Figure 1. Categorizing AI failure significance and AI autonomy

[1] Multiple definitions of automation and autonomy exist; however, it is important to have a clear understanding of the differences between automation and AI-enabled autonomy. Automation is considered to be a system that automatically acts on a specific task according to pre-defined, prescriptive rules. For example, reactor protection systems are automatically actuated when process parameters exceed certain defined limits. Autonomy is a set of intelligence-based capabilities that allows the system to respond to situations that were not pre-programmed or anticipated (i.e., decision-based responses) prior to system deployment. AI-enabled autonomous systems have a degree of self-governance and self-directed behaviour resulting in the ability to compensate for system failures without external intervention.

As the amount of AI autonomy increases along the horizontal axis, there is a tendency to move from AI that informs human decision-making to AI that conducts operations with little or no human oversight. Between the levels of Insight/Collaboration and Operation/Full Autonomy, a transition exists where there is potentially less time for humans to respond to AI faults as the AI begins to function less as a design, engineering, or recommendation tool, and more as tool which replaces human decision-making. As the significance of AI failure increases along the vertical axis, there is a potential for greater scrutiny of the overall AI system and how it impacts safety and security. If the AI system functions in region one, there may be greater flexibility in deployment. Developers and evaluators of AI systems may need to consider the corresponding impact and consequences for AI systems that function in other regions. This chart is one way to consider the overall AI system risk and does not make claims about how such risk can be quantified.

Existing international standards which cover the development of digital components and systems in nuclear safety applications (e.g. IEEE 7-4.3.2 and IEC 60880) provide requirements on the use of software tools to develop software-based systems performing safety functions.[2] These existing standards may well be considered applicable for the AI categories shown above in Figure 1, or at least considered as a starting point when developing AI systems. The remainder of this section will discuss the significance of AI failure in regions three and four compared to regions one and two.

Failure that Could Significantly Impact Safety or Security (Regions Three and Four)

AI systems which function in region three are characterised by the ability to verify the system's output before it is actioned. However, while there may be increased time to respond to AI system failures, an unrevealed error could significantly impact safety, so a robust verification process is required for any outputs. Use cases falling under this region may include systems that aid in the design or the maintenance of safety systems. As with existing safety systems, there remains the potential for common-cause failure in AI systems that are provisioned across, for example, multiple reactor facilities.

Conversely, AI systems which function in region four are characterised by a lesser ability to verify the system's output before it is actioned. While it may be technically possible to verify the output using humans, the key distinction between three and four is that this verification may not be possible within a reasonable time limit. Region four AI systems, therefore, require the AI system to function properly, or alternatively for other more robust components or systems to mitigate potential failure of the AI. Use cases falling under this region may include AI-optimised control and protection algorithms.

Failure that has Minimal or No Impact on Safety or Security (Regions One and Two)

If the risk associated with AI system failure is sufficiently low, it may have minimal or no direct impact on safety and security. However, AI could impact safety as a secondary consideration that may not be highlighted through existing engineering practices, as in the following example.

[2] The World Nuclear Association developed a crosswalk document entitled "International Nuclear I&C and Electrical System Standards Tables with URLs" which brings together the nuclear power plant instrumentation and control (I&C) and electrical system standards from the Institute of Electrical and Electronics Engineers (IEEE) and International Electrotechnical Commission (IEC). This document provides a useful and condensed comparison of IEEE and IEC standards for a variety of nuclear specific application areas.

An AI system is deployed on a nuclear reactor turbine plant to analyse maintenance data and determine whether certain maintenance items can be reduced. This AI system could introduce a reduction in plant maintenance that increases the likelihood of spurious turbine trips followed by subsequent reactor trips, placing undue demand on the reactor safety systems. It is a key safety principle that reactor protection systems are not unduly stressed.

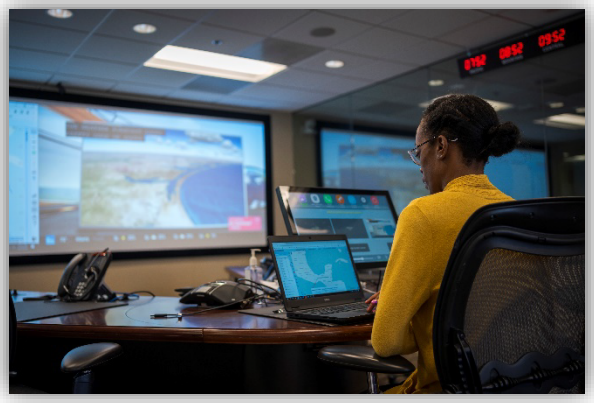
In this example it becomes clear that increased risk could result unexpectedly from AI deployment. As with all risk, this will have to be managed proportionately by the nuclear operator. AI systems which are used, for instance, to make operational decisions or direct decision making are therefore more reliant on ensuring the underlying AI model is updated, maintained, and qualified. As AI implementation transitions from region one to two, the AI system will need to be maintained to avoid performance degradation and be kept consistent with the pre-determined change control and notification process for that application. Region two usage increases the level of AI-enabled autonomy and potentially decreases the available response time for human-led actions to intervene with AI faults.

Considering whether the tool has the capability to introduce a fault into the software system, whether it can miss an existing fault, and whether the output of the tool can be verified through other means will influence how the AI system or component is categorized under the relevant safety classification scheme. Even if the AI system is determined to be non-safety related and lower system confidence is deemed acceptable, it remains important to consider that fault tolerance, avoidance, and software verification are important aspects of digital computer system design. Section 4 on the use of existing safety and security systems engineering principles details this idea further.

4 USE OF EXISTING SAFETY AND SECURITY SYSTEMS ENGINEERING PRINCIPLES

Regulators require nuclear dutyholders or licensees to demonstrate that risks arising from nuclear activities are adequately controlled and to articulate how these risks have been managed. Safety principles for a variety of nuclear activities are codified in international standards and country-specific regulatory guidance for the engineering of safety and security systems aimed at addressing these expectations. These principles provide direction for how systems that could cause harm are designed and are well established in the nuclear industry. However, these standards and guidance were not written with AI technologies in mind, so AI developers and system integrators will have to consider which parts can be applied and what more should be done to address any gaps.

Engineering standards state that the most appropriate technology should be used for a given application. This technology should be as simple as possible to give assurance that failure types can be analysed and understood, with as few unknown failure types as possible. Therefore, in many applications, an effective approach to managing risks arising from AI system failures may be the deployment of the simplest possible technology in conjunction with AI to prevent unknown failures from causing harm. In short, AI may not be the appropriate solution as it could unnecessarily complicate an otherwise straightforward system. Simpler technologies generally reduce the overall system uncertainty and lead to greater assurance the design objectives have been achieved.



For example, a nuclear operator wishes to deploy an AI system in the control room of a nuclear power plant to aid reactor operator decision-making. There are no specific standards relating to this, but it would be reasonable for a regulator to expect that the nuclear operator would identify the most appropriate standards, regulations, and regulatory guidance that could be applied. In this situation, there are standards and guidance that cover control rooms, functional software safety, categorisation of functions and classification of systems, and so on. A gap analysis could highlight areas where standards and guidance cannot be applied due to unique aspects of the AI system, enabling the nuclear operator to identify appropriate mitigations for managing the residual risk. These mitigations could be AI-based techniques, to gain some confidence in how the AI could fail, or conventional systems to which existing standards fully apply. Table 1 in the Annex gives a view of typical standards and guidance that could be applied in the above example across the nuclear domain.

In certain applications, the consequences of AI failure may be so high that it would be very challenging to justify its use. Additionally, there may be other use case areas where conventional technologies cannot intervene to effectively prevent the consequences of AI failure. In time, design and analysis techniques may be developed to verify AI can be relied upon where significant consequences arise from failure. These verification models would need to be clearly demonstrated to enable deployment of AI in these applications.

Conversely, for certain applications the consequences of AI failure may be tolerable, particularly if the safety benefit provided by the AI demonstrably outweighs the risk of failure. For example, in the case of removing

operators from hazardous processes in a glove box, the risk of the AI damaging containment may be negligible compared to the benefit of reduced dose to operators. In this case, it would be advisable for system designers to consider recovery from failure conditions to account for all sources of risk associated with AI failure.

Other principles beyond simplicity have been proven to be effective in managing risk, including diversity, redundancy, separation, and segregation. These measures encompass an approach to designing and operating nuclear facilities that prevents and mitigates accidents that release radiation or hazardous materials. The key is creating multiple independent and redundant layers of defence to compensate for potential failures so that no single layer, no matter how robust, is exclusively relied upon. These approaches are likely to remain applicable for AI systems, such that safety will not be wholly dependent on any single element of the design, construction, maintenance, or operation of AI in nuclear applications.

5 HUMAN AND ORGANISATIONAL FACTORS

AI technologies enable increasingly autonomous systems. Even though AI is a potential technology that could be used to achieve autonomy, it may be possible to have autonomy without relying on AI. Additionally, not all uses of AI are fully autonomous. For instance, many AI capabilities may be used to augment human decision-making rather than replace it. Because of this blending of machine and human decision-making, organisations and end users may be presented with unique human and organisational factor challenges. This section considers that while humans may deploy AI systems with a clear objective, it may be difficult to determine if this objective has ultimately been met. Adding a machine decision maker also means that the human may or may not be responsible for the behaviour of the machine. This can have profound impacts on areas including reactor operator licensing, concepts of operation, and rulemaking. Therefore, consideration should be given to how AI failure impacts human-machine teaming, in view of the potential of increasing reliance on AI-driven decision-making and more human-like machine interaction (e.g. conversational AI assistants).

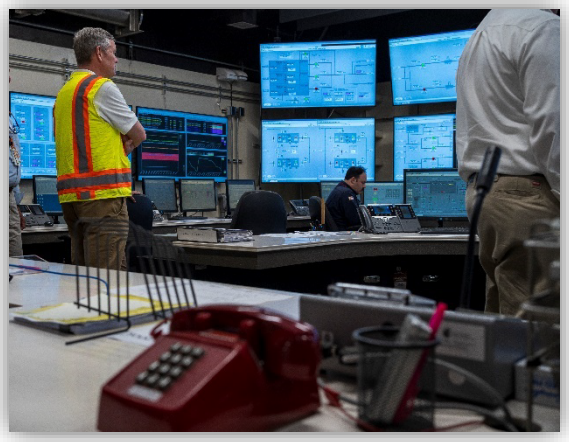
Existing regulatory frameworks at the CNSC, UK ONR and US NRC outline how nuclear workers should be trained, certified, verified, and subsequently monitored to confirm they are performing as required. Over the decades, human capabilities and vulnerabilities have been well established. A foundational component of human factors engineering is developing human performance baselines which capture quantitative and qualitative characteristics of the computer system within the overall nuclear control system (IEEE Standards Association, 1998[3]). As AI will impact both quantitative and qualitative aspects of a nuclear system, it is good practice to thoroughly describe its intended use to support subsequent analyses, design decisions, and compliance programs. Considerations in this area may include, but are not limited to, the following:

- the capability the AI component or element is intended to provide,
- which functions, roles, or responsibilities will be allocated to the AI as opposed to humans,
- novel functions, roles, or responsibilities arising from the inclusion of the AI capability,
- how humans interact with the AI, and
- humans' ability to intervene, or not, with the operation of the AI (level of autonomy).

[3] IEEE Standards Association (1998) IEEE Guide for Information Technology - System Definition - Concept of Operations (ConOps) Document, Retrieved from IEEE Standards Association: https://ieeexplore.ieee.org/document/761853

Human factors practitioners have long used various means to consider how functions can best be distributed between humans and machines, including for enabling functions such as work schedules, training, and staffing provisions. The criteria used for the allocation of functions to humans or machines, including AI, may consider the relative capabilities and weaknesses of current technologies. Additionally, whenever functions are allocated to a machine, sufficient regard should be given to related supervisory or monitoring tasks allocated to a human. This is particularly important for AI capabilities, given the current level of uncertainty in how to reliably identify limitations or faults.

Many modern AI systems act as black boxes, meaning little information is available to the user regarding how the system generates an output from its inputs. This is notably different than most nuclear applications where, for example, reactor operators can memorize the logic circuits and can readily identify when the circuit is not working as expected. Reactor operators are currently trained to trust their control panel indications (e.g. meters, gauges, annunciators, displays, etc.) because these are reliable, predictable, and unexpected operation is usually easy to discern (such as through channel checks or by using alternate instruments). Humans may be reluctant to trust an AI system and may second guess any decision or action made autonomously. Alternatively, humans may completely trust the AI system, assuming the machine is always correct, meaning human verification is effectively bypassed. This could lead to complacency where the human does not monitor the system because it usually functions well and thus the human is no longer an effective backup to the AI system. Preventive guardrails such as more traditional software systems or enhanced training and protocols could be deployed to reduce risk.



Effective integration of AI into nuclear systems will need to explore the continuum of trust between the human and machine, establishing an optimal level that will realize the benefits of AI while maintaining appropriate human oversight. In establishing an appropriate level of trust between the human and the AI capability, attention should be given to:

- the level of system knowledge required of the human,
- how humans monitor the performance of the AI capability,
- how the AI capability self-monitors,
- how the need for human intervention is identified, either by the human or the AI capability itself,
- how the handover of control between the AI capability and human is accomplished, and
- if a human has intervened to take control of functionality from the AI capability, when and how control can be returned to the AI.

In addition to assessing the above items during the development of AI capability for a nuclear system, thought should be given to managing trust and guarding against human complacency throughout the life of the system. New features indicating uncertainty associated with the performance and alarms to indicate questionable performance may be required.

All system functions allocated to humans require a user interface containing all the relevant information presented in an effective way. AI capabilities may introduce novel information or require innovative means of data presentation to allow humans to effectively manage the system functions while maintaining an appropriate level of understanding and trust with the AI. Additionally, user interfaces for AI systems should consider how to support the transfer of control between the human and AI in both normal and failure conditions.

A training program should be developed whenever a new technology or system is introduced, and AI in nuclear systems is no exception. In addition to considering the knowledge required to understand the system's operations and AI capabilities, training programs may need to address methods for identifying AI failures, processes, and procedures for intervention (e.g. taking control from or correcting the AI), and processes and procedures for restoring the AI capability in the system. As AI systems can evolve while in service, thought should also be given to structuring training programs in such a way that they can be adapted with any evolution in AI capability. AI capabilities could even be introduced in the development or delivery of training programs.

Finally, thought should be given to the implications of introducing AI capabilities on human performance and organizational monitoring programs – for example, the potential impacts on safety culture. During the development phase, it should be possible to verify that safety has been established as the top priority in decision-making algorithms. Once deployed, however, how will decisions made by AI models be monitored and validated to confirm they are consistent with the intended safety priorities? Consideration should also be given to how findings from safety culture monitoring will be fed back into the system design, from adjustments to the AI models through to the re-allocation of functions from the AI to the human. Additionally, consideration should be given to maintaining a minimum level of personnel qualification for jobs that are being augmented by AI and supporting appropriate human redundancy, training, and credentialing to ensure organizational effectiveness.

The introduction of AI capabilities could significantly change how humans interact with nuclear systems. Careful thought should be given prior to introduction of an AI system on how to effectively incorporate AI while maintaining the appropriate level of human and organizational performance to ensure safe operation in a nuclear environment.

6 AI ARCHITECTURE IN NUCLEAR APPLICATIONS

Integrating AI into nuclear systems requires careful consideration of software and control system architectural principles to ensure reliable and secure operation. This section discusses some considerations around AI architecture in nuclear applications such as system boundaries, monitoring, and modularization. Architecture, as discussed in this section, refers to how AI may interact with the wider system and is likely to be a powerful consideration in managing the deployment of safe and secure AI. It should be noted that principles utilized in areas of software and system architecture for nuclear applications could still be applicable and should be encouraged.

System boundaries

Establishing clear system boundaries is crucial for ensuring the integrity and security of AI-enabled nuclear applications. These boundaries define the scope of the AI system, delineating its interaction with the physical nuclear environment and other components if the AI system influences control or protection systems. System boundaries should be thoughtfully designed to ensure data exchange is reliable, secure, and compatible with the existing infrastructure. Additionally, these boundaries should incorporate mechanisms for handling errors and exceptions, preventing the AI system from inadvertently triggering unsafe actions, or disrupting critical processes. Some system boundary considerations include:

- Data availability – AI can work with a large variety of data sources, which if unrestricted can lead to inefficiency within system processing. As systems are limited through hardware with a set processing power, boundaries can be determined early on to optimize resources.

- Constraining software input and output – Bounding software systems by placing limits on software input and output that are controlled by conventional systems may allow trust to be placed in the wider system architecture rather than the AI component itself. If the allowable range of AI outputs are safe, there may be a lower likelihood of undesirable or unintended consequences.

- Isolation principles – The overall system architecture may be used to bound the AI system functionality. Diverse, redundant, and isolated systems are a key safety and security principle for nuclear facilities, and isolation from stakeholder or interfacing system control can minimize unintended actions.

Monitoring

Continuous monitoring of AI systems is essential to detect potential anomalies, prevent failures, and maintain system integrity. Monitoring strategies should encompass both the AI system itself and the physical nuclear environment.

AI system monitoring focuses on assessing the performance, accuracy, and reliability of the AI models and algorithms. This may involve tracking metrics such as model accuracy, error rates, computational efficiency, and bias. For example, monitoring may include a mechanism for detecting and alerting to potential adversarial attacks

or attempts to manipulate the AI system. This may involve implementing anomaly detection algorithms, intrusion detection systems, and access control measures.

An effective monitoring setup within the overall architecture may ensure accuracy in AI outputs and build trust within the AI system. The concept of monitoring plays hand-in-hand with boundaries, as boundaries can limit what would need to be monitored.

Modularization

While monolithic AI finds trends and consumes data from a variety of sources, AI used in nuclear applications may benefit from modularization to promote flexibility, maintainability, and explainability. Modular design principles involve dividing the AI system into smaller, independent modules, each with a well-defined function. Regarding modularization, software and data are particular facets to be considered.

From the software perspective, there are a variety of AI methods and techniques (e.g. computer vision, natural language processing and generative language models) that are widely applied across a variety of use cases. Because of this variety of use cases, AI systems developed for non-safety applications may not have been held to the same rigorous development standards that are required in safety applications. Thus, domain adaptation may be a concern where the AI model training and validation differences from one domain to another could lead to subtle omissions or errors resulting from incomplete AI knowledge.

Transferring AI models into new domain applications will require careful consideration of the risks, benefits, and operational performance differences. For example, it may be appealing to repurpose a modular computer vision application trained and deployed to monitor fire in a non-safety auxiliary building to then monitor and detect fire in a safety-related equipment room. However, consideration should be given to the new operational conditions, model limitations, and performance characteristics in the new environment if it was not part of the original AI system development.

The modularity of an AI system, and thus its applicability in other areas (e.g. transferring a computer vision weld inspection model used on one type of metal weld to another type of metal weld in a different operational environment), is highly dependent on the data used to train, validate, and test that model. Understanding the bounding conditions of how the model was trained and the associated application limitations should be considered when evaluating a new application use case and accounted for in the overall risk-informed decision-making.

Modularity is potentially beneficial for the overall system design because it allows issues or errors to be isolated at the level of the problematic module without adversely affecting the functionality of the larger operational system. Thus, modularity benefits both the developer's and end user's ability to track performance and trace back the root cause of faults. Large, monolithic AI models may result in AI systems whose behaviour is difficult to interrogate, explain, or diagnose.

7 AI LIFECYCLE MANAGEMENT

This section discusses high level principles for AI lifecycle management in nuclear applications, as well as how to understand and manage the AI lifecycle as an iterative process from design conception to development and finally deployment. These AI lifecycle considerations include critical attributes of AI lifecycle management, highlighting the need for evaluation processes, configuration management strategies, and ongoing monitoring and adaptation. This is necessary to mitigate risks and ensure safe and secure integration of AI applications.

Considerations for AI lifecycle management

Interdisciplinary teams are beneficial for managing the AI lifecycle effectively, including design, development, and deployment as part of the overall product lifecycle. Some unique considerations when combining traditional nuclear safety and AI lifecycles are discussed below.

- AI design – Data is the foundation of any AI solution, so specifying the data required and its structure is crucial.

- AI development – The model selection and training process is iterative, and fine-tuning will likely be required to achieve the performance objectives for a given u se case. Appropriate quantifiable metrics should be selected to communicate model performance and accuracy to stakeholders.

- AI deployment – A model deployed in a production environment may be exposed to new data not represented by the training data. Therefore, consideration should be given to implementing automated continuous monitoring controls that can detect and mitigate data and model drift. Following deployment, systems should be established to monitor for unexpected behaviour that may be a result of training data bias or model failure.

While traditional software engineering lifecycle management practices should be maintained, there are a few unique AI-specific lifecycle management considerations useful to factor in when evaluating the entire development, maintenance, and monitoring approach.

- Testing, evaluation, verification, and validation – Tools and metrics are needed to assess a model's ability to perform as intended while adhering to performance, security, and ethical standards. The topic of AI ethics encompasses a broad range of topics where human values are imbued on AI systems such as individual rights, privacy, non-discrimination, and humane decision-making. Continuous monitoring of deployed AI models is essential to ensure their continued performance and adherence to established criteria. This involves tracking metrics, as well as identifying and addressing potential performance degradation.

- Configuration management and rapid technology change – AI systems may require more frequent maintenance and triggers for conducting corrective maintenance due to data drift, where the original data is no longer representative of the operational data, and model drift, where the original model no longer faithfully represents the underlying phenomena being modelled. Both concepts lead to performance degradation or

introduction of biases. AI research, development and available tools are constantly updating, with this pace of change intensified by the open-source nature of available packages.

- Managing risks from retraining, domain adaptation and data drift – Datasets used to train AI systems may become detached from their original and intended context or may become stale or outdated relative to deployment context. AI models may need to be retrained to maintain their effectiveness in response to changing data distributions or evolving business requirements. However, retraining poses risks such as overfitting or introducing new biases. Careful consideration of retraining strategies, including the selection of appropriate datasets and evaluation techniques, is crucial to mitigate these risks. Further, concerns exist around domain adaptation of AI models trained on specific datasets, which may not perform optimally when applied to different domains or environments. It is important to consider risks associated with third-party AI technologies; transfer learning, where knowledge learned from one task is re-used on another related task; and off-label use, where AI systems may be trained for decision-making in one domain and then fine-tuned for another. AI models may also exhibit performance discrepancies between test environments and real-world deployments due to factors such as data biases, hardware differences, and environmental conditions.[4]

Relevant AI lifecycle management standards and practices

In the published literature, as well as the AI guidance and frameworks published by numerous national entities, there is no shortage of different AI lifecycle models, frameworks, and process concepts. Many of these AI lifecycle management approaches, whether domestically or internationally recognized, attempt to extend existing software and system lifecycle processes and many have not yet reached the level of a final consensus standard.[5]

In the absence of well-established AI-specific standards, non-AI specific nuclear software and system lifecycle processes likely remain broadly applicable to AI systems. In using these standards, the unique attributes introduced by AI should be considered. In addition, there are a myriad of guides attempting to bridge the gap between academic research and practical commercial implementation of AI product design. These guides can be found in academic literature, AI vendor best practices and guidance, and national rubrics for developing safe and secure AI systems. They include aspects of design, development, and deployment such as user needs definition, data collection, mental models, explainability and trust, feedback and control, and errors and graceful failure.

[4] US National Institute of Standards and Technology (NIST), AI Risk Management Framework, Appendix B, https://airc.nist.gov/AI_RMF_Knowledge_Base/AI_RMF/Appendices/Appendix_B

[5] A collection of standards and guidance across regulatory areas is provided as an Annex to this document. The list contains both existing non-AI and AI-specific standards and guidance relevant to the nuclear industry.

8 DOCUMENTING AI SAFETY AND SECURITY

AI has been used in specific scientific applications for many decades, but it has only recently been made practical in a wider range of applications through improvements in low-cost computing power and tools. The nuclear industry is typically a slow adopter of new technology, as has been experienced with obsolescence management and conversion of instrumentation and controls from analog to digital hardware. Therefore, there is currently limited experience with the deployment of AI systems in nuclear safety and security applications and instead much ongoing research on potential research and development use cases.

AI systems have the potential to fail in ways both similar to and different from conventional systems and humans. The causes of failure are likely to be difficult to determine because it is not currently feasible to transparently interpret how the output of an AI system relates to its inputs. Many existing verification and validation activities applied to conventional software during the design process may be inadequate to quantify all the errors in AI systems. Also, because many AI systems are designed to adapt to evolving data and change the way they behave over time in response to feedback, previous performance confirmation can quickly become obsolete.

The consequences of incorrect output will vary according to the application, with some applications being more sensitive than others. This could result in a range of different outcomes, from no effect to unacceptable risks to people, the facility, or the environment. It is important to document the basis for AI safety and security in a manner that enables common understanding by a variety of stakeholders including users, developers, and regulators. Currently no method exists to quantify the failure probability of an AI component within a system, which makes it difficult to trust AI components to perform a function with any level of integrity. As discussed above, this integrity may have to be derived from the architecture of the AI system and its components. There are however certain aspects of the AI systems themselves that need to be understood if claims are to be placed on other system components to reduce the risk of AI failure.

The way in which an AI system may fail should be understood to demonstrate overall system safety. An AI system may fail in unexpected ways, and this failure may propagate in unexpected ways. This could significantly alter traditional failure analysis if not appropriately taken into account. However, if components are used to limit the possible outputs of AI, it may be possible to eliminate the consequences of unexpected AI failure. Similarly, there are many different types of AI technologies, and it can be challenging to demonstrate that a technology will always achieve a specified performance outside of a laboratory environment. The uncertainty of specific AI technologies needs to be clearly understood. In the absence of certainty, consideration should be given to the tolerability of consequences of failure or their management by another independent system of sufficient reliability.

Data, testing, and security will all likely be challenges. Unique considerations will be needed to verify the adequacy of data for AI systems and how bias, uncertainties, and data artifacts may affect the demonstration of safety. It is advisable to ensure the failure of the AI is tolerable or mitigated by more robust systems. Testing is one element commonly used to provide confidence in system performance under a range of conditions. However, as systems become more complex, the number of internal states rapidly multiplies, and adequate

testing becomes increasingly difficult. Most AI systems would increase complexity to an extreme level, making it impossible to gain any significant assurance of an AI system's safety from testing alone.

Demonstrating the security of AI systems may require novel approaches as there are likely novel ways in which AI behaviour can be altered by malicious activity, such as using the data a system is trained on to target it. The attack vector space is likely to be even greater than that of conventional software systems due to the increase in complexity.[6]

Finally, documentation will be important for presenting the case that AI systems are adequately safe. Aspects of this documentation will be similar to existing systems such as verification and validation manuals, testing procedures, methods, and user guides.

[6] The security risk posed to nuclear facilities due to AI cyber intrusion techniques is out of scope for this paper.

9 CONCLUSION

The nuclear industry benefits from decades of operational experience, mature and rigorous design and operation protocols, and a strong safety and security culture. The rapid pace of recent AI development is somewhat antithetical to the slow and methodical change process that the nuclear industry traditionally follows. Nevertheless, the primary goal for the nuclear industry and regulators with respect to AI systems will be maintaining adequate safety and security while benefiting from their deployment. Some concluding points to emphasize include thoughtfully addressing security challenges in AI systems, maintaining a perennial focus on data, and being mindful of consensus standards.

Securing AI is challenging as many AI systems contain components from external or open sources such as data, software (including operating systems), hardware, and hardware configurations. The principles described in this paper relate to both safety and security considerations. The security challenges for AI systems are at least as significant as for existing software approaches even if AI systems are subject to novel security vulnerabilities that need to be considered alongside standard cyber security threats. Guidelines for secure AI system development were formulated in cooperation with 22 agencies and ministries from across the world, including security agencies from the UK, US, and Canada. The document recommends guidelines for any systems that use AI, whether those systems have been created from scratch or built on top of tools and services provided by others. Implementing these guidelines will help providers build AI systems that function as intended, are available when needed, and work without revealing sensitive data to unauthorised parties.[7]

Data is a foundational component underpinning the development and implementation of AI systems. It serves as the catalyst for AI algorithms to learn, identify patterns, and make informed predictions. In the nuclear industry, where reliability, safety, and security are paramount, the quality and integrity of data are critical for ensuring adequate performance of an AI system. Deficiencies in data, such as incompleteness, bias, or inaccuracies, could lead to erroneous AI outputs with potentially unacceptable consequences; flawed data may result in flawed output. Equally, AI systems receiving very different inputs in real-world deployment versus training or testing may not perform well. Therefore, meticulous data curation, validation, and governance across the AI lifecycle is central for the successful integration of AI into the nuclear industry.

Engineering design standards help establish activities and good practices that support key objectives including safety and security. International standards exist for conventional software, hardware, and systems development in the nuclear sector. However, standards for nuclear-specific AI design and substantiation do not currently provide sufficient assurance in high consequence applications or where societal harms might exist. The fast pace of AI development means it is unlikely that AI-specific consensus standards for the nuclear domain will be available to support regulatory activities within the near future. In the interim, existing nuclear-specific standards remain a starting point coupled with considering the unique attributes introduced by AI.

[7] Further information on “Guidelines for Secure AI System Development” can be found at https://www.ncsc.gov.uk/files/Guidelines-for-secure-AI-system-development.pdf.

While there are hurdles to consider to successfully deploy AI, there are also potentially significant benefits to using AI. If effectively managed, negative consequences could be avoided or mitigated for many applications. This document recognizes this position and describes features the Canadian, UK, and US nuclear regulators consider important in managing risks arising from the use of AI.

FOR THE CANADIAN NUCLEAR SAFETY COMMISSION:

BY: 

NAME: Keith Dewar

TITLE: Director Innovation and Research

DATE: 16 August 2024

PLACE: Ottawa, Ontario Canada

FOR THE UNITED KINGDOM OFFICE FOR NUCLEAR REGULATION:

BY: 

NAME: Shane Turner

TITLE: Technical Director

DATE: 27 August 2024

PLACE: Liverpool, United Kingdom

FOR THE UNITED STATES NUCLEAR REGULATORY COMMISSION:

BY: 

NAME: John R. Tappert

TITLE: Acting Director, Office of Nuclear Regulatory Research

DATE: 8/15/24

PLACE: Rockville, Maryland United States of America

10 FURTHER READING

This section provides a list of links to useful documents, repositories, standards, websites, and organizations that further expand upon some of the considerations discussed in this document surrounding the safe and secure development and usage of AI. Not all links are nuclear-specific, but many are generally applicable to a broad range of AI use cases considered in the nuclear domain. Reference to any specific standards, commercial products, processes, or services does not constitute or imply its endorsement or recommendation on the part of the Canadian Nuclear Safety Commission (CNSC), United Kingdom Office for Nuclear Regulation (UK ONR), or the United States Nuclear Regulatory Commission (US NRC).

- The Organisation for Economic Co-operation and Development (OECD) AI Policy Observatory maintains a repository of tools and metrics to help AI actors develop, deploy, and use trustworthy AI systems and applications.
- The AI Standards Hub is a UK initiative dedicated to the evolving and international field of standardisation for AI technologies and curates a searchable database of international AI standards.
- The US NRC's AI Strategic Plan, covering fiscal years 2023–2027, establishes the vision and goals for the US NRC to cultivate an AI-proficient workforce, keep pace with AI technological innovations, and ensure the safe and secure use of AI in US NRC-regulated activities.
- The US National Institute of Standards and Technology (NIST) developed the NIST AI Risk Management Framework (AI RMF) intended for voluntary use and to improve the ability to incorporate trustworthiness considerations into the design, development, use, and evaluation of AI products, services, and systems.
- The US Department of Energy developed the AI Risk Management Playbook as a comprehensive reference guide for AI risk identification and recommended mitigations (actionable pathways) to support responsible and trustworthy (R&T) AI use and development.
- The US Food and Drug Administration (FDA), Health Canada, and the United Kingdom's Medicines and Healthcare products Regulatory Agency (MHRA) jointly identified 10 guiding principles that can inform the development of Good Machine Learning Practice.
- The US National Security Agency's Artificial Intelligence Security Center (NSA AISC) published the joint Cybersecurity Information Sheet Deploying AI Systems Securely in collaboration with CISA, the Federal Bureau of Investigation (FBI), the Australian Signals Directorate's Australian Cyber Security Centre (ASD ACSC), the Canadian Centre for Cyber Security (CCCS), the New Zealand National Cyber Security Centre (NCSC-NZ), and the United Kingdom's National Cyber Security Centre (NCSC-UK).
- MITRE Adversarial Threat Landscape for AI Systems (ATLAS™) is a globally accessible, living knowledge base of adversary tactics and techniques based on real world attack observations and realistic demonstrations from AI red teams and security groups.
- AI Watch is the artificial intelligence website of the European Commission's Joint Research Centre (JRC), which presents the outputs of activities in Trustworthy AI.

- The Institute for Ethical AI & Machine Learning maintains a GitHub repository which contains a curated list of open-source libraries and commercial platforms which could be used to deploy, monitor, version and scale artificial intelligence and machine learning applications.
- The Google People + AI Guidebook provides a set of methods, best practices, and examples for AI design.
- The AI Incident Database curates a repository of user submitted incident reports aimed at indexing the collective history of harms or near harms realized in the real world by the deployment of artificial intelligence systems.
- The International Organization for Standardization (ISO) and the International Electromechanical Commission (IEC)'s joint technical committee (JTC) 1, subcommittee 42 (JTC1/SC42) focuses on AI standards.
- The Institute of Electrical and Electronics Engineers (IEEE) developed the IEEE CertifAIEd™ certification program for assessing ethics of Autonomous Intelligent Systems to help protect, differentiate, and grow product adoption.
- The IEEE Artificial Intelligence Standards Committee has a number of approved and proposed AI-related standards across a variety of technical domains.

11 ANNEX

Table 1: Relevant standards and guidance across regulatory areas

Topic	Standards and guidance documents
General	REGDOC 2.5.2 Design of Reactor Facilities US NRC, 10 CFR Part 50, Domestic licensing of production and utilization facilities US NRC, 10 CFR Part 50, Licenses, Certifications, and Approvals for Nuclear Power Plants UK Nuclear Installations Act 1965 UK ONR, Safety Assessment Principles for Nuclear Facilities, 2014 Edition, Revision 1
System safety series	IEEE, 603, Standard Criteria for Safety Systems for Nuclear Power Generating Stations (Note: This standard is based on single failure criterion. It is not considered as “functional safety” standard.) IEC 61513 Nuclear power plants – Instrumentation and control important to safety – General requirements for systems & IEC 61508 Functional safety of electrical/electronic/programmable electronic safety-related systems
Functional safety	IEC 61513 Nuclear power plants – Instrumentation and control important to safety – General requirements for systems IEC 61508 Functional safety of electrical/electronic/programmable electronic safety-related systems IEEE 603™-2018, Standard criteria for safety systems IEEE 279™-1971, Criteria for protection systems (Note: Both IEC 61508 and IEC 61513 are considered as "functional safety" standards. IEC 61513 is a "daughter" standard of IEC 61508 specifically applicable to nuclear industry.)

Topic	Standards and guidance documents
Software safety	IEC 60880 Nuclear Power Plants – Instrumentation and Control Systems Important to Safety – Software Aspects for Computer-Based Systems Performing Category A Functions IEC 62138, Nuclear Power Plants – Instrumentation and Control Important for Safety – Software Aspects for Computer-Based Systems Performing Category B or C Functions IEEE, 7-4.3.2, Standard Criteria for Digital Computers in Safety Systems of Nuclear Power Generating Stations IEC 62566, Nuclear power plants - Instrumentation and control important to safety - Development of HDL-programmed integrated circuits for systems performing category A functions. IEC 62566-2 Nuclear power plants - Instrumentation and control systems important to safety - Development of HDL-programmed integrated circuits - Part 2: HDL-programmed integrated circuits for systems performing category B or C functions. IEEE 1012™-2016 or 2004, System and Software Verification and Validation IEEE 1028™-2008, Software reviews and audits IEEE 828™-2012 or 2005, Configuration management systems and software engineering IEEE 829™-2008, Software and system test documentation IEEE 1008™-1987, Software unit testing IEEE 830™-1998, Software requirements specifications
Hardware safety	IEC 60987 Nuclear Power Plants – Instrumentation and Control Important to Safety – Hardware Design Requirements for Computer-Based Systems IEEE 7-4.3.2™-2016 or 2003, Criteria for programmable digital devices in safety systems

Topic	Standards and guidance documents
Cyber security	CSA N290.7:21 Cyber Security for Nuclear Facilities CNSC REGDOC 2.5.2, Design of Reactor Facilities CSA Group, *Emerging Digital Technologies Within the Canadian Nuclear Industry*, https://www.csagroup.org/article/research/emerging-digital-technologies-within-the-canadian-nuclear-industry/ *Guidelines for secure AI system development* (Canada, US, UK and international partners): https://www.cyber.gc.ca/en/news-events/guidelines-secure-ai-system-development International Atomic Energy Agency, *Computer Security Techniques for Nuclear Facilities*, IAEA Nuclear Security Series No. 17-T (Rev. 1), IAEA, Vienna (2021). International Atomic Energy Agency, *Computer Security of Instrumentation and Control Systems at Nuclear Facilities*, IAEA Nuclear Security Series No. 33-T, IAEA, Vienna (2018). International Atomic Energy Agency, *Computer Security for Nuclear Security*, IAEA Nuclear Security Series No. 42-G, IAEA, Vienna (2021). International Atomic Energy Agency, *Computer Security Approaches to Reduce Cyber Risks in the Nuclear Supply Chain*, Non-serial Publications, IAEA, Vienna (2022). IEC 62645:2019 - Nuclear Power Plants - Instrumentation, control, and electrical power systems - Cybersecurity requirements, IEC, Geneva, 2019. IEC 63096:2020 - Nuclear power plants - Instrumentation, control and electrical power systems - Security controls, IEC, Geneva, 2020. IEC 62443 Series of standards for security for industrial automation and control systems IEC TR 63486, Nuclear Facilities - Instrumentation, control and electrical power systems - Cybersecurity risk management approaches IEEE 692-2013 - IEEE Standard for Criteria for Security Systems for Nuclear Power Generating Stations
Control room series	IEC 61839:2000, Design of control rooms - Functional analysis and assignment IEC 61771:1995, Main control-room - Verification and validation of design IEC 61772:2009, Control rooms - Application of visual display units (VDUs) NUREG-0700, R2, 2002, Human-System Interface Design Review Guidelines

Topic	Standards and guidance documents
Categorisation of functions and classification of systems	IEC 61226 Nuclear power plants - Instrumentation, control and electrical power systems important to safety - Categorization of functions and classification of systems IEC/TR 61838 Nuclear power plants -Instrumentation and control important to safety - Use of probabilistic safety assessment for the classification of functions IEC 62385 Nuclear power plants - Instrumentation and control important to safety - Methods for assessing the performance of safety system instrument channels IEC 1500 Nuclear power plants – Instrumentation and control important to safety, Data communication in systems performing category A functions IEEE 1819™-2016, Risk-informed categorization of electrical and electronic equipment
Lifecycle management	ISO/IEC FDIS 5338[8], *Information technology — Artificial intelligence — AI system lifecycle processes,* defines a set of processes and associated concepts for describing the lifecycle of AI systems based on machine learning and heuristic systems ISO/IEC FDIS 5338 provides processes that support the definition, control, management, execution, and improvement of the AI system in its life cycle stages[9] IEEE 1074™-2006, Developing a software project lifecycle process

[8] ISO/IEC FDIS 5338 is based on ISO/IEC/IEEE 15288:2023 and ISO/IEC/IEEE 12207:2017 with modifications and additions of AI-specific processes from ISO/IEC 22989:2022, *Information technology - Artificial intelligence - Artificial intelligence concepts and terminology*, and ISO/IEC 23053:2022, *Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML) – First Edition*.

[9] When an element of an AI system is traditional software or a traditional system, the software lifecycle processes in ISO/IEC/IEEE 12207:2017 and the system life cycle processes in ISO/IEC/IEEE 15288:2015 can be used to implement that element.