

Vienna Center for Disarmament and Non-Proliferation



April 2025

Past, Present, and Future Applications of AI in the Nuclear Sector

Donald Dudenhoeffer

Vienna Center for Disarmament and Non-Proliferation | vcdnp.org

Authors



Donald D. Dudenhoeffer is a Senior Cyber Security Consultant with over 35 years of experience in the nuclear and security fields. Currently, he is supporting the commercial nuclear sector, working as an Executive Consultant at the Barakah Nuclear Power Plant in Abu Dhabi, UAE as well as serving

as a part-time researcher at AIT Austrian Institute for Technology. Previously, Mr. Dudenhoeffer was a Senior Information Technology Officer for the Division of Nuclear Security (NSNS) at the International Atomic Energy Agency (IAEA), where he led the Computer Security programme, providing Member States with guidance and support in developing security programmes for their nuclear facilities.

He holds a Master of Science degree in Operations Research from the US Naval Postgraduate School and has previously qualified as a nuclear engineer.

About the VCDNP

The Vienna Center for Disarmament and Non-Proliferation (VCDNP) promotes international peace and security by conducting research, facilitating dialogue, and building capacity on nuclear non-proliferation and disarmament.

The VCDNP is an international non-governmental organisation, established in 2010 by the Federal Ministry for European and International Affairs of Austria and the James Martin Center for Nonproliferation Studies at the Middlebury Institute of International Studies at Monterey.

Our research and analysis provide policy recommendations for decision-makers. We host public events and facilitate constructive, resultsoriented dialogue among governments, multilateral institutions, and civil society. Through in-person courses and online resources on nuclear nonproliferation and disarmament, we train diplomats and practitioners working in Vienna and around the world.

Acknowledgements

This research and paper were made possible through the support of the Vienna Center for Disarmament and Non-Proliferation (VCDNP) as well as a research project funded by **Global Affairs Canada**. Additionally, the author would like to thank the survey contributors for their insightful perspectives and comments.



Sponsored by

Contents

Foreword by DiplIng. Helmut Leopold, PhD i
Foreword by Elena K. Sokova iii
Al Overview
AI Applications – State of Practice in Nuclear 7
Nuclear Industry Perspective
Demographics
Drivers and Use
Concerns and Barriers
Survey Conclusions
Al Implementation Challenges
Data
Infrastructure Requirements
Workforce Impact
Trustworthiness: Transparency and Explainability 30
Dependability
Malicious Use and Exploitation
Al Governance Overview
Laws of Robotics
Standards for AI Use and Development 41
Nuclear Regulation on Al Application 43
Conclusions

Foreword



The Responsible Use of Artificial Intelligence Needs a New System Understanding

by Dipl.-Ing. Helmut Leopold, PhD

Head of the Center for Digital Safety & Security AIT Austrian Institute of Technology

Driven by the rapid digitalisation in all spheres of our lives, data science and artificial intelligence (AI) have today become a cornerstone for innovation, problem-solving, and business transformation. Al-based business opportunities and solutions for the essential challenges faced by society, the environment, and healthcare are enabled and driven by modern data science approaches. Data, whether structured or unstructured, numerical, textual, or audiovisual, put in context with other data, analysed and processed by smart algorithms, are the basis for smart solutions for industry, for eGovernment, and even for our broad private usage of digital services. Integrating these data flows, more and more AI systems are being used in a wide variety of application areas because they consistently achieve better results than conventional approaches. The hype surrounding the impressive performance of AI goes hand in hand with the availability of extremely powerful and inexpensive processors and storage capacity as well as affordable IT-services offered on a global scale. Thus, there is also a strong promise of AI to address the challenges in the nuclear sector and promote nuclear energy, nuclear disarmament, and nuclear non-proliferation goals.

However, as AI becomes increasingly popular, we must develop new methodologies for system design and operation, as well as methods for testing and certification of AI systems, to ensure a system behaviour which meets our ethical principles, laws, regulations, and safety performance requirements. With respect to nuclear, when data science and AI are proposed for use in critical safety and security environments, we urgently need methods to build trust in and assurance of AI system operations. While tasks may be AI-enhanced, we must retain ultimate responsibility for what the machines do. The value of human-centric decision-making and control must still be maintained.

Thus, digital transformation requires a level of technological mastery for the harmonious integration of AI into industry and, more broadly, society. If our society does not face up to this emerging challenge, we could create a situation of unchecked technology integration with potential adverse impacts across many areas of our lives. Adverse events will likely include acts by malicious actors possessing novel and enhanced AI-assisted capabilities for their objectives. In this light, we must find mechanisms to increase the robustness of our society as well as to ensure the sustainable control of AI systems while establishing the conditions for the beneficial use of AI within society. An essential component of this goal is to establish research initiatives in order to better recognise and understand the new risks associated with AI technologies and to develop suitable countermeasures.

To protect against the improper use of AI, it is imperative to understand these new technologies better, to define standards, and implement policies and regulations for its socially acceptable use, as well as to develop skills and tools, so that authorities, companies, and even individual citizens have the opportunity to participate in AI's positive developments to society as a responsible media and technology consumer.

The nuclear industry encapsulates an enormous, interconnected supply chain that circles the globe. Digital transformation, including AI integration, has the potential to provide great benefits in improving efficiencies, reducing costs, and improving operational safety.



For the reasons previously discussed, caution must prevail, however. Nuclear security, specifically, preventing unauthorised access to nuclear materials or the occurrence of a sabotage event remain key drivers when considering the use of AI. There must be a balance between AI integration and the risk it might impose.

To build awareness and explore the use of AI in the nuclear industry, the Vienna Center for Disarmament and Non-Proliferation (VCDNP) has established a close cooperation with the Austrian Institute of Technology, the largest Austrian research and technology organisation. The goal of this collaboration is to foster a sustainable interdisciplinary dialog in the nuclear sector for awareness-raising and education and as a basis for responsible AI technology development and deployment.

The following report gives an in-depth insight into the development of AI in general and the application in security and safety critical application domains, to build the basis for subsequent critical dialogue on building and using next generation AI systems in the nuclear industry.

Dipl.-Ing. Helmut Leopold, PhD And Mill



Foreword



Artificial Intelligence and the Nuclear Sector

by Elena K. Sokova

Executive Director Vienna Center for Disarmament and Non-Proliferation (VCDNP)

Modern AI systems are increasingly being considered for a range of applications in nuclear facilities, not only in new advanced reactor designs but also for integration into currently operating facilities. These systems could provide significant benefits in applications, such as perimeter monitoring, preventative maintenance, and facial recognition, among others. However, these systems also pose new challenges for facilities that use and store nuclear materials. The nuclear industry, policy-makers, and regulators must arrive at these benefits through mitigating the novel nuclear security risks that could be associated with these systems.

To understand the challenges AI systems and models integrated into nuclear facilities present, it is important for all relevant stakeholders to improve their AI literacy. Further, it is important to have reliable processes in place to assess the quality, robustness, and transparency of these AI systems and models. Questions need to be posed – and answered – regarding the types of applications AI systems are being integrated into and any associated vulnerabilities that could be exploited by a malicious actor.

The current paper was produced as part of a research project conducted by the Vienna Center for Disarmament and Non-Proliferation (VCDNP), supported by Global Affairs Canada, on the intersection of AI, nuclear security, and the security of the nuclear supply chain. This paper provides a needed overview of the evolution of AI in the nuclear sector as well as current and future applications of AI in this sector. The paper also provides analysis of a survey of the nuclear industry on the use of AI, which provides useful insights into industry thinking and preparedness on integrating AI applications into nuclear facilities. It also suggests some key questions to consider when seeking to integrate AI systems and models into nuclear facilities. The paper is well suited to serve as a basis for stakeholders to understand the connection between AI and the nuclear sector now and into the future.

The collaboration on the research project and this paper allowed for a close cooperation between the VCDNP and the Austrian Institute of Technology, to inform and raise awareness of relevant stakeholders about the rapidly shifting changes in technology and its applications in the nuclear sector. The digital transformation in all sectors, but especially the nuclear sector, requires a multistakeholder approach where both technical and policy experts are needed to find the right balance between the benefits of technology integration and the associated security risks.

Elena K. Sokova





While artificial intelligence is not a new concept, the technology has made significant leaps in recent years, which have unlocked new capabilities and an unprecedentedly wide range of applications.

AI Overview

Beginnings

The 1950s are generally considered the naissance of the science of artificial intelligence (Al). While previous scientific and fictional works had postulated thinking machines and robots, the 1950s saw the seminal work of Alan Turing, John McCarthy, and an array of others who brought forward the concept of artificial intelligence as a potential reality and a field of research.

During this same period, other significant advances in science and engineering were occurring in the nuclear field. In 1951, the world saw the first use of nuclear fission to produce electricity as the Experimental Breeder Reactor 1 produced electricity to light the town of Arco, Idaho. In 1954, the United States Navy launched the first nuclear-powered submarine, the USS Nautilus. That same year, the Soviet Union completed the Obninsk Nuclear Power Plant, the first grid-connected nuclear reactor. The United States followed shortly after in 1957 with the commissioning of the Shippingport Atomic Power Station in Pennsylvania.

The past over 70 years have seen dramatic journeys for both AI and nuclear power. Both have evolved and matured, albeit both having to overcome significant challenges. The nuclear industry was originally seen as the source of endless electrical power, but three significant accidents (Three Mile Island, Chernobyl, and Fukushima) changed the nuclear landscape and led to the downfall of many nuclear facilities, darkening the future of nuclear energy development. Climate change and the insatiable need for electric power by computing technologies, such as AI, however, have brought forward the beginnings of a nuclear renaissance with nuclear power increasingly being considered in future energy portfolios.



The application of AI into mainstream use, while not suffering from the consequences of catastrophic events, such as Chernobyl, has likewise had a slow path.¹ Despite 70 years of research and development, AI has often failed to live up to the expectations surrounding it. The last 20 years, however, have seen tremendous breakthroughs in AI applications, due in part to advancements and economies of computational power and access to enormous quantities of data, which can be used to train AI models.

Today, AI is seen as an essential enabler by businesses and industry, including the nuclear power industry. AI applications, such as Microsoft's Copilot and Google's Gemini, are already becoming more mainstream in day-today business and social uses, while other AI technologies are being considered for use in operations and maintenance applications.

This paper provides a practical overview of AI technologies and examines the application of AI technologies in nuclear facilities, namely in nuclear power plants. The goal of this paper is not to provide a taxonomy of technologies, but to examine the state of practice, lessons learned, barriers, and industry perceptions related to AI integration into facility work processes.

Collection of Technologies

What is AI? The term "artificial intelligence" or "AI" is being used as a catalyst for great innovation and transformation, but what exactly does AI entail? Conducting an online search of the definition of "artificial intelligence" will result in thousands of responses. Here is the definition of Artificial Intelligence from Microsoft's Copilot (AI providing a self-definition):

Artificial intelligence, or AI, refers to the simulation of human intelligence in machines that are programmed to think, learn, and problem-solve. These systems can process information, recognize patterns, make decisions, and adapt to new situations. Essentially, it's like teaching machines to think and act like humans. Cool stuff, right?²

The term "artificial Intelligence" was first coined in 1956 at the Dartmouth Summer Research Project on Artificial Intelligence by Professor John McCarthy to identify machines which could perform tasks that are characteristic of human intelligence.³ This original framing of the term has withstood the test of time.

This leads us to the next question: "What are the characteristics of human intelligence?" Examples in this case might include visual perception, speech recognition, and decision-making, among others. Computational processes have, in recent years, made significant achievements in each of the above domains resulting in a range of practical applications for industry. However, while advances have been made, there are limitations associated with AI that are important to understand. The achieved magnitude of "intelligence" by an AI is broadly classified into three types:⁴

• Narrow (or Weak) AI: Narrow AI is trained to perform a single or "narrow" task, often far faster and better than a human mind can. However, it cannot perform outside of its defined task. It targets a single subset of cognitive abilities and advances in that spectrum.

2 Definition from Query to CoPilot "Definition of Artificial Intelligence" on 13 October 2024. Available at: <u>https://copilot.microsoft.com/?sendquery=1&FORM=SCCODX&showconv=1</u>.

4 IBM, "Understanding the different types of artificial intelligence", 12 October 2023. Available at: <u>https://www.ibm.com/think/topics/artificial-intelligence-types</u>.



¹ Here, one must understand that AI is not a single entity, but a collection of tools and technologies developed to promote a level of autonomous analysis, decision-making, and now, generative ability.

³ Calum McClelland, "The Difference between Artificial Intelligence, Machine Learning, and Deep Learning", medium.com, 4 December 2017. Available at: https://medium.com/iotforall/the-difference-between-artificial-intelligence-machine-learning-and-deep-learning-3aa67bff5991.

- Artificial General Intelligence (AGI) (or Strong AI): AGI uses previous learning and skills to accomplish new tasks in a different context without the need for human beings to train the underlying models. This ability allows AGI to learn and perform any intellectual task that a human being can. AGI remains theoretical at present.
- **Super AI:** Super AI is commonly referred to as true artificial superintelligence and, like AGI, is strictly theoretical. If ever realised, Super AI would think, reason, learn, make judgements, and possess cognitive abilities that surpass those of human beings.

AI Techniques

It is important to understand that AI is not a single technology, but a collection of data science tools and techniques, which are constructed to provide specific aspects of "intelligence", such as problem-solving, learning, or recognition. These tools and techniques, among others, consist of knowledge-based systems, machine learning, neural networks, and deep learning.

Knowledge-based Systems: Knowledge-based (or expert) systems consist of a programme, which encapsulates knowledge from a specific domain, normally obtained from a human expert in that domain.

Machine Learning (ML): As a subset of computer science, ML leverages data and algorithms to mimic human learning processes. By continuously analysing and learning from training data, ML models enhance their accuracy over time. This field encompasses various techniques, each offering unique approaches to problem-solving and data analysis.

Neural Networks (NN): NNs are a technique used in machine learning. Neural networks consist of node layers, containing an input layer, one or more hidden layers, and an output layer. Each node connects to another and has an associated weight and threshold. If the output of any individual node is above the specified threshold value, that node is activated, sending data to the next layer of the network. Otherwise, no data is passed along to the next layer of the network by that node. Through training, the weights and thresholds are adjusted such that the input maps to a desired output. NNs are commonly used for pattern recognition and classification tasks.

Deep Learning: Deep Learning is a subset of NNs. The "deep" in deep learning refers to the number of layers in a neural network. A neural network that consists of more than three layers – which would be inclusive of the input and the output – can be considered a deep learning algorithm or a deep neural network (DNN). In other words, an NN containing several hidden layers would be considered a DNN. A neural network that only has three layers (input layer – hidden layer – output layer) is a basic (not deep) neural network. Figure 1 illustrates a Deep NN.



Fig. 1: (Deep) Neural Network with six layers



Predictive and Generative Models

The application of AI technologies, especially in industry, has traditionally been in the application of predictive models. The predictive models seek to analyse and solve computationally complex and data-intensive problems. Through the analysis of patterns of historical and live data, such systems have been developed to support equipment performance analysis, maintenance planning, work process optimisation, and decision support.

Generative models are designed not only to analyse and learn from massive data streams, but they are designed to "create" new or derived content from their learning and user specification. Large Language Models (LLMs) have been developed to support text generation, image generation, video generation, and so forth. Figure 2 represents a rendering by an AI artist using Microsoft Copilot and the prompt "create an image of an artist creating an image". To date, generative AI models, such as Microsoft's Copilot and OpenAI's Chat GPT-4, have become a common component in day-to-day business processes, though significant adoption of these models in industrial facilities remains less frequent.



Fig. 2: Generated by Copilot using the prompt [create an image of an artist creating an image]

AI System

For practical industrial applications, various techniques, including those mentioned above, are combined to develop an "AI System" equipped with computer vision, natural language processing, document intelligence, automation, and/or decision support capabilities, tailored to solve specific organisational challenges or perform specific work processes. A nominal AI system is shown in Figure 3.

In this case, a Deep Neural Network is the core technique used to train the AI system. One of the first steps in model creation consists of data collection and training. This training effectively creates the AI model that will process or react to data input. Data input can originate from various sources, including sensors in a vision system or user prompts in a chatbot. Based upon the input, the AI system will process the information consistent with its learning to produce an associated output or action. Note that many AI systems incorporate several AI technologies. For example, a chatbot system may employ technologies, such as natural language processing and document intelligence, to both interact with a person and uploaded documents. A more detailed description of the AI lifecycle is provided later in the paper.





Fig. 3: AI System Overview (adapted from OECD)⁵

Enablers

As mentioned previously, the study of AI technologies and their use is not new, and despite successes in limited applications, has – more times than not – failed to live up to its promise. The repeated failure to meet expectations led to periods of so-called "AI Winters", in the 1970s and1980s, and again from the 1990s through the 2000s, characterised by a lack of interest and funding in AI research.⁶ What has changed in the last decade to renew promise in AI applications, introducing the beginning of the recent "AI Spring"?

Many reasons exist for the current AI boom, but there are three key factors that have driven its recent advancement:

- High-quality data and data access
- Al models or algorithms
- Computational resources

First, **high-quality data** is the foundation of any AI learning-based model. High-quality, accurate, non-biased, and non-corrupted data is required for both the training of the AI model, but also for its execution. Gaining access to such data therefore is essential for building an effective and reliable model. Trends, such as the digitalisation of the world, the growing Internet of Things, and the emergence of Industry 4.0, have led to the proliferation of data collection from an ever-expanding network of digital devices. This collection capability as well as the associated data highway(s)⁷ created are much better aligned today than they were in decades past to support the creation and collection of critical data sets to support a growing array of AI applications. This is most certainly true in the nuclear industry, where ten years ago, the digital footprint of a nuclear power plant (NPP) may have been a thousand "digital assets", made up of computers, Programmable Logic Controllers (PLCs), network appliances, and so forth. Today, the number of digital assets can easily exceed ten times that amount in traditional NPP construction units. This digital reliance is expected to continue to grow with advanced reactor designs. Each digital asset in the NPP has the potential to collect plant and equipment performance data that could be injected into an AI model.

⁷ A data highway refers to high-speed and high-capacity data networks that enable the rapid flow of information between entities. Put more simply, it is like a high-speed network connectivity supporting rapid data access and transmission.



⁵ Organisation for Economic Co-operation and Development, "Artificial Intelligence

in Society", 11 June 2019, p. 23. Available at: https://www.oecd.org/en/publications/artificial-intelligence-in-society_eedfee77-en.html.

⁶ More background on the history of AI winters can be found on Wikipedia, "AI winter", available at: https://en.wikipedia.org/wiki/AI_winter.

Second, **the AI model or algorithm** consists of the underlying computer programme(s) supporting both learning and execution of the model. As previously stated, AI is a collection of techniques. The past ten years have seen great advancements in the use of deep learning for AI applications in industry. Further, the tailoring of AI algorithms to meet specific performance tasks has demonstrably advanced. The development of these applications is not confined to the research laboratory, as in decades past, but has blossomed into a billion-dollar industry, with both commercial and open-source frameworks available enabling collaboration and innovation for a greater population of consumers.

Finally, **computational resources** are greater than in the past. AI models are computationally intensive, which previously posed a consequential barrier to their effective use in industry because the computing power available at the time was not sufficient to create timely results. One of the most significant AI hardware developments occurred in the 2010s with the advancement and specialisation of graphics processing units (GPU). Originally designed for providing high quality graphics, these units are also well suited to support the parallel processing associated with training deep learning models. GPUs have now been optimised for AI applications. Google has also developed specialised hardware, the tensor processing unit (TPU), which is an AI accelerator application-specific integrated circuit (ASIC), developed to support neural network machine learning with the Google TensorFlow software.⁸ Cloud computing has also provided tremendous benefits for the advancement and proliferation of AI use. Cloud computing can provide the necessary infrastructure for an organisation to implement AI without the requirement to create their own internal hardware and software infrastructure.

8 Google Cloud, "Accelerate AI development with Google Cloud TPUs", Cloud Tensor Processing Units. Available at: https://cloud.google.com/tpu?hl=en#accelerate-ai-development-with-google-cloud-tpus.





While predictive AI has seen limited use in nuclear power plants so far, the opportunities for generative AI use remain underexplored.

AI Applications - State of Practice in Nuclear

The following section discusses the current and potential use of AI in the nuclear industry. While many organisations have already implemented AI as part of standard business processes, i.e., Copilot, Gemini, and ChatGPT, this section looks at AI applications in nuclear power plant processes. In the nuclear industry, predictive AI has long been proposed and implemented in a limited fashion within the realm of performance monitoring and maintenance support. Today, there is an interest in not only leveraging advances in predictive AI, but also generative AI, such as LLMs which provide a natural language interface to support human interaction. However, LLMs, like Copilot, Gemini, and ChatGPT, require tremendous quality training data sets as well as computing and network infrastructure to be practically implemented.

For the nuclear industry, instead of using an LLM to draft a business report to summarise the notes of a meeting, imagine using an LLM to generate a synopsis of the Operator Experience surrounding an infrequent and highly consequential maintenance task. As another example, imagine using an LLM to automatically parse a cyber security advisory, retrieve relevant equipment lists of susceptible items, and report any indicators of compromise. LLMs have raised the potential for practical application in nuclear. This section will provide a high-level discussion of potential use cases for both predictive and generative AI.



Nuclear Operations

The proposed application of AI technologies in nuclear facilities has been ongoing since the late 1980s. At that time, researchers examined the application of neural networks, expert systems, fuzzy logic, and genetic algorithms and their potential to enhance the safety, reliability, and operability of nuclear power plant systems. Robert Uhrig, for example, one of the pioneers in the use of neural networks in industrial applications, proposed AI technologies for applications, such as transient identification, plant-wide monitoring, analysis of vibrations, and performance, condition, and efficiency monitoring.⁹ Since then, the landscape has drastically changed. The digital transformation and the drive to increase operational efficiencies, improve system visibility, and enhance safety have driven the nuclear industry to explore AI as an essential component for future investment.

In a notable example, operation and maintenance (O&M) costs represent between 40 and 70 percent of overall generating costs, with a significant portion devoted solely to labour expenses. Digital transformations, including the integration of AI technologies, are being evaluated to increase plant efficiencies while reducing O&M costs. Significant research in this respect considers using AI technologies for process optimisation, equipment monitoring, and equipment maintenance. Reduction of human task requirements, especially in hazardous domains, has also been the focus of developments in robotics for potential use in nuclear facilities. Recent advancements in AI and supporting technologies has further broadened the scope of assessment for AI integration.

Much research has been done on various applications of AI systems in nuclear power plants. Table 1 provides a non-exclusive list of both researched and implemented applications of AI technologies in nuclear operations. Tang,¹⁰ Lu,¹¹ and Huang¹² provide more detailed reviews of the associated research in the application of AI technologies at nuclear facilities.

9 Robert E. Uhrig, J. Wesley Hines, and William R. Nelson. "Integration of artificial intelligence systems for nuclear power plant surveillance and diagnostics," Short Course Notes, 1998. Available at: <u>https://scholar.google.ae/scholar?</u>

<u>q=hines+%22Integration+of+Artificial+Intelligence+Systems+for+Nuclear+Power+Plant+Surveillance+and+Diagnostics%22&hl=en&as_sdt=0_ &as_vis=1&oi=scholart.</u>

11 Chao Lu et al., "Nuclear Power Plants With Artificial Intelligence in Industry 4.0 Era: Top-Level Design and Current Applications—A Systemic Review", IEEE Access, Vol. 8, 20 October 2020, pp. 194315-194332, DOI: 10.1109/ACCESS.2020.3032529.

12 Qingyu Huang et al., "A review of the application of artificial intelligence to nuclear reactors: Where we are and what's next", Heliyon, Volume 9, Issue 3, 2023. Available at: <u>https://doi.org/10.1016/j.heliyon.2023.e13883</u>.



¹⁰ Chenwei Tang et al., "Deep Learning in Nuclear Industry: A Survey", Big Data Mining and Analytics, Vol. 5, Issue 2, June 2022, pp. 140-160, DOI: 10.26599/BDMA.2021.9020027.

Design and Engineering	Reactor core designThermal-hydraulics simulation and analysesRadiation shield design	
Manufacturing/Supply Chain	Defect detection (e.g., nuclear fuel assemblies, fuel pellets)Detection of counterfeit, fraudulent, and suspect items (CFSI)	
Operations and Maintenance (O&M)	 Online condition monitoring Fault detection and diagnosis Real-time defect detection Predictive maintenance Process automation Inspection and assessment (robotics/Uncrewed Aerial Vehicles (UAV)) Mobile monitoring (robotics) Data/log collection (robotics/UAV) Work planning/schedule optimisation Report automation/generation 	
Decommissioning	Dismantling (robotics)Hazardous area survey (robotics/UAV)	
Disaster/Event Response	 Firefighting (robotics) Remote surveillance/atmospheric sampling (robotics/UAV) Search and rescue (robotics/UAV) 	

Table 1: Application of AI Technologies in Nuclear Operations

Digital Twins

A digital twin is the virtual representation of an object or system designed to reflect a physical object accurately.¹³ Key attributes often include real-time data feeds and machine learning to produce realistic status updates and projections on performance. Digital twins are not themselves an AI system but often incorporate AI technologies to represent system dynamics and predict system performance.

Simulations and digital twins are similar in nature, but differ in two primary ways. Digital twins nominally rely upon real-time data feeds from the physical system that they are emulating, whereas this is not a common feature in many simulations. Another difference, although not as substantial, is that of scope. Digital twins typically provide a comprehensive representation of a system, while simulations are often more narrowly focused on what they are trying to examine. Thus, digital twins are well suited for large physical or complex projects, for which situational awareness is needed but may not be readily evident from current data feeds and modelling. Digital twins can be used to enhance the assessment of system performance, rapid identification of potential faults, and better inform decisions regarding operations, maintenance, and event response.

Interest in the digital twin concept is growing in the nuclear industry. In the United States, the US Nuclear Regulatory Commission (USNRC) has conducted research along with Idaho National Laboratory (INL) and Oak Ridge National Laboratory (ORNL) regarding the development and regulatory viability of a digital twin.¹⁴ Figure 4 presents a view of their vision.

13 IBM, "What is a digital twin?", 5 August 2021. Available at: <u>https://www.ibm.com/topics/what-is-a-digital-twin</u>.
14 US Nuclear Regulatory Commission (USNRC), "Digital Twins", 2023. Available at: <u>https://www.nrc.gov/reactors/power/digital-twins.html</u>.





Fig. 4: USNRC Concept of Digital Twin Application in Nuclear¹⁵

The implementation of a digital twin for a nuclear facility offers significant benefits for both licensees and regulators. Digital twin technology could provide a shared platform for real-time situational awareness, thereby enhancing safety margins and regulatory efficiencies. Such a platform can lead to regulatory advantages, including the immediate identification of safety issues, increased responsiveness, and more effective oversight.¹⁶

Nuclear Safety

Nuclear safety, as defined by the International Atomic Energy Agency (IAEA), is "the achievement of proper operating conditions, prevention of accidents and mitigation of accident consequences, resulting in protection of workers, the public and the environment from undue radiation risks."¹⁷ Safety analysis broadly examines natural hazards, material fatigue, equipment failure, and human error. Within the nuclear facility, safety systems are designed to minimise the risk and impact of an adverse event by maintaining reactivity control and system integrity, thus reducing undue radioactive exposure to the environment or people. The operation of safety systems is driven by deterministic and ensured behaviours. However, as observed by Helmut Leopold, "Conventional deterministic technical systems can be comprehensively tested, with their behaviours traced and analysed by experts, to ensure safe operation. This is vital for safety-critical systems. As this is not possible for AI systems, we need new approaches for effectively testing AI systems, as well as understanding and explaining decision-making processes."¹⁸

15 Ibid.

18 Helmut Leopold, "Mastering Trustful Artificial Intelligence", in: Rene Schmidpeter & Reinhard Altenburger (eds.), Responsible Artificial Intelligence, CSR, Sustainability, Ethics & Governance, Springer, pp. 133-158, February 2023. Available at: https://doi.org/10.1007/978-3-031-09245-9_6.



¹⁶ Raj Iyengar, Official Transcripts of the Proceedings, Nuclear Regulatory Commission, "Advisory Committee on Reactor Safeguards", 4 May 2022, p. 96. Available at: <u>https://www.nrc.gov/docs/ML2217/ML22179A369.pdf</u>.

¹⁷ International Atomic Energy Agency, "IAEA Nuclear Safety and Security Glossary", 2022 (Interim) Ed., p. 139. Available at: https://www-pub.iaea.org/MTCD/Publications/PDF/IAEA-NSS-GLOweb.pdf.

Al explainability standards and accepted testing processes have not been fully adopted yet, which makes the application of Al in direct support of nuclear safety systems unlikely in the near term. Additionally, the lack of nuclear regulations addressing the use of Al applications in nuclear facilities and the caution of facility operators to implement Al solutions over existing methods in the current fleet of nuclear power plants makes it likely that the nuclear industry will see only limited application of Al in plant operations in the coming years.

While today's nuclear power plants encompass an array of automated safety features, plant operation and control are still very dependent on human action. Unfortunately, statistics have shown that approximately 80 percent of events at nuclear facilities have been either caused by or adversely impacted by human performance.¹⁹ With this in mind, there is increased interest in using AI applications to enhance or support human performance and reduce human error.



Fig. 5: Performance Shaping Factors (adapted from Blackman et al. (2008))

Research has shown that human performance can be modified either positively or negatively by "shaping factors" related to the individual, the task, and the environment. Figure 5 illustrates a taxonomy of human performance shaping factors.²⁰ Thus, well-designed AI systems, that support human perception, cognition, and response, may have the potential to positively shape human performance. Research efforts are underway to evaluate AI applications for decision support and human error reduction in industrial and nuclear facilities.^{21,22}

Notably, AI applications that reduce the complexity of a situation, reduce workload, and positively guide operator action could be of value. Likewise, AI applications that assess human performance for errors or estimate fitness for duty may be able to mitigate and correct potential adverse human actions. Table 2 provides a list of possible areas for AI integration.

https://www.standards.doe.gov/standards-documents/1000/1028-BHdbk-2009-v1/@@images/file.

21 Mohsen Soori et al., "Al-Based Decision Support Systems in Industry 4.0, A Review", Journal of Economy and Technology, 28 August 2024. Available at: <u>https://doi.org/10.1016/j.ject.2024.08.005</u>.

22 Ezgi Gursel et al., "Using artificial intelligence to detect human errors in nuclear power plants: A case in operation and maintenance", Nuclear Engineering and Technology, Vol. 55, Issue 2, February 2023, pp. 603-622. Available at: <u>https://doi.org/10.1016/j.net.2022.10.032</u>.



¹⁹ US Department of Energy, "DOE Standard: Human Performance Improvement Handbook", Volume 1: Concepts and Principles, DOE-HDBK-1028-2009, June 2009, pp. 1-10. Available at :

²⁰ Harold S. Blackman, David I. Gertman, and Ronald L. Boring, "Human Error Quantification Using Performance Shaping Factors in the SPAR-H Method", Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Vol. 52, Issue 21, September 2008, DOI: <u>10.1177/154193120805202109</u>.

Human Augmentation and Decision Support	 System analysis and performance estimation Alarm management and prioritisation Human error/anomalous action detection Plant status assessment and future projection (situational awareness support) Complex information processing and situation assessment "Intelligent" Human Machine Interface (HMI) "Intelligent" Operating Procedures Operator task augmentation Abnormal event assessment and decision support Al-driven training for customised learning Operator experience/technical document review and retrieval Operator fitness assessment (i.e., capability to fulfil task)
--	--

Table 2: Application of AI Technologies for Human Support

AI Dependency

While AI is an extremely powerful tool, which can support an array of operations at a nuclear facility, as shown in Table 1, dependency on AI technologies can result in potential adverse consequences, especially with regards to human performance.

Through the utilisation of AI technologies, humans may become passive rather than active participants in certain activities. For example, humans may allocate key functions to AI, such as situation assessment, intuitive analysis, critical thinking, and creative problem-solving. Automated processes could also potentially supplant previous staff task-based skills. This leads to several essential questions that need to be addressed. If a dependency on AI develops, how should necessary human-centric capabilities and skills be maintained? Additionally, how are humans re-integrated into the decision or task process if the AI capability is removed?

As an illustration, consider the simple example of a car with autonomous driving features, such as autonomous emergency braking, adaptive cruise control, blind spot detection, and high-definition near-real time maps. These features most certainly enhance driver safety and the driving experience, but how does a driver react if such capabilities are suddenly lost, and does a dependence on such features reduce the ability to operate the vehicle in a non-autonomous/non-supported mode?

Research on autonomous vehicle technology has identified multiple potential impacts on human performance, suggesting that, with the integration of autonomous features in a car, the human driver can experience:²³

- Possible loss of driving skill
- Loss or reduced state or awareness of system processes
- Greater difficulty in troubleshooting errors
- Slower response in assuming and retaking control during an abnormal event or loss/limitation of automation

At present, human operations are still central to nuclear plant operations and security operations. The introduction of AI and autonomous functions into a nuclear facility will need to not only be effective but also need to assess human impact and recovery from a loss of such functions.

23 Adela Spulber and Richard Wallace, "Impact of Automated Vehicle Technologies on Driver Skills", Michigan Department of Transportation and The Center for Automotive Research, 30 June 2016. Available at: <u>https://www.cargroup.org/publication/impact-of-automated-vehicle-technologies-on-driver-skills/</u>.



Nuclear Security

Nuclear security, like nuclear safety, seeks to prevent undue radiological risk to the environment and people. However, rather than focusing on accidents, nuclear security focuses on the prevention of, detection of, and response to, criminal or intentional unauthorised acts involving or directed at nuclear material, other radioactive material, associated facilities, or associated activities.²⁴ These criminal or intentional unauthorised acts may include physical and/or cyber attacks. Prevention, detection, and response measures for both physical security and cyber security include physical, technical, and administrative control measures. Al has applications to support/automate the technical aspects of prevention, detection, and mitigation of criminal or intentional unauthorised acts, often referred to simply as "malicious acts". When Al is integrated into nuclear security systems in this way, the discussion related to Al human performance support for nuclear safety also applies to nuclear security.

Physical Security

A shorthand often used when discussing physical security of nuclear facilities²⁵ is "guns, guards, and gates". It is therefore easy to question what AI has to do with physical security. To answer this question, it is important to understand the basic principles of action for physical security at a nuclear facility which are: deter, detect, delay, respond, and eliminate. The goal of the physical protection system (PPS) is to prevent a malicious act or to interrupt such an act before it is accomplished. The timeline for an adversary attack and the PPS response is visualised in Figure 6.



Fig. 6: Adversary Timeline Overview

It is not assumed that AI could replace the human guards or the response force, but AI could support and enhance their timely response and ultimate success. AI systems could, for example, support the PPS to better prevent initial access, decrease detection and assessment time (t_d) , and reduce response and interdiction time (t_i) . Examples of current AI applications to support physical security are listed in Table 3, which includes biometric access control readers, video analytics that assist in detecting/identifying anomalous behaviour, and alarm/early-notification management.

24 International Atomic Energy Agency, "Objective and Essential Elements of a State's Nuclear Security Regime", Nuclear Security Series, No. 20, 2013. Available at: <u>https://www.iaea.org/publications/10353/objective-and-essential-elements-of-a-states-nuclear-security-regime</u>.

25 Physical security in this application refers to physical protection, security operations, and physical security functions.



AI Technologies in Physical Protection	 Access Control – Identify/Verify Facial recognition Fingerprint scanning Voice recognition Alarm assessment Alarm management Anomaly detection (e.g., to combat insider threat) Change/tamper detection Data fusion and correlation (enhanced situational awareness) Physical intrusion detection Predictive analysis and decision support (PPS design, response planning) Remote/autonomous monitoring (robots/drones) Routine task automation Threat identification – trustworthiness checks, background investigations Threat intelligence
	 Threat identification – trustworthiness checks, background investigations Threat intelligence Video surveillance analytics

Table 3: Applications of AI Technologies in Physical Protection

Cyber Security

Cyber security has become very data-driven in the use of end-point security applications,²⁶ network monitoring, as well as the implementation of Security Operations Centres (SOC). Digital assets and associated systems are increasingly being actively monitored and assessed against identified vulnerabilities and known indicators of compromise. Anomalous behaviour is investigated for indications of compromise or potential insider threat activity.

Further, and importantly, the supply chain has become a potential vector of compromise. New digital systems, new digital assets, software applications, software patches, and associated integrations must be assessed for supply chain exploits, introduced vulnerabilities, and system performance impact. These tasks, among others in cyber security, have seen the explosion of data and complex system analysis requirements. Many of these tasks, however, may be well suited for the application of AI, and in fact, AI has already been integrated into many commercial cyber security products, such as user behaviour analytics (UBA), intrusion detection systems (IDS), intrusion prevention systems (IPS), and data loss prevention (DLP) systems. In these cases, AI supports the analysis of massive data sets and the identification of abnormal or restricted human behaviours as well as system behaviours indicative of potential malicious activity.

The application of AI in cyber security will continue to grow, addressing the increasing complexity of digital systems and the vast amounts of data they generate. Similarly, threat actors have already adopted AI technologies to carry out malicious activities. As these attacks become more sophisticated, AI will be essential for detecting and defending against them. Table 4 provides a list of current and near-term production applications of AI in the cyber security field. Additional discussion of ongoing research and the state of practice is provided in "A Survey of Artificial Intelligence in cyber security" by Moravat.²⁷

²⁷ Katanosh Morovat and Brajendra Panda, "A Survey of Artificial Intelligence in cyber security", 2020 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 2020, pp. 109-115, DOI: 10.1109/CSCI51800.2020.00026.



²⁶ End-point security applications are designed to provide prevention, detection, and mitigation capabilities to components connected to the network, such as computers, smart phones, servers, etc. Nominally, such applications provide for logging and data extraction, which provides a source of analytical data.

Table 4: Application of AI Technologies in Cyber Security

28 A botnet is a collection of often compromised internet devices that can be used to conduct a distributed denial of service (DDoS) attack against a target.





Survey data on the nuclear industry reveals key insights about readiness of industry to integrate AI into operations.

Nuclear Industry Perspective

While the promise of AI is high and many uses for nuclear facilities have been proposed, what is the perspective of the industry at present, and what is the state of practice? To better understand this perspective, the author conducted a Nuclear Industry Survey to compile thoughts and practices on AI adoption in nuclear facilities from a range of stakeholders. The results of the Nuclear Industry Survey are compared to the UK Department for Science, Innovation & Technology AI Cyber Security Survey, conducted by IFF Research,²⁹ and other recent surveys to identify any noted differences in perspectives among the nuclear sector compared to other industries, including drivers for use and potential barriers.

Demographics

The survey was sent to nuclear stakeholders active in national and international organisations, such as the IAEA, the World Institute of Nuclear Security (WINS), and the Nuclear Threat Initiative (NTI). These stakeholders may be more aligned with nuclear security functions and especially cyber security than other nuclear power plant aspects. This is relevant, however, in that the cyber security group will often be the function trusted to conduct associated security and risk analyses of new digital technologies prior to acceptance.

29 UK Government, Department for Science, Innovation & Technology, "AI cyber security survey – main report", 15 May 2024. Available at: https://www.gov.uk/government/publications/research-on-the-cyber-security-of-ai/ai-cyber-security-survey-main-report.



Organisations

The nuclear industry consists of an array of stakeholders that included but are not limited to competent authorities, licensees (dutyholders), researchers, suppliers, vendors, and so forth. The survey looked to capture a sampling of views on AI across these areas. Figure 7 shows the distribution of survey participants.



Fig. 7: Participant Organisations

Participant Roles

The participants, as noted below in Figure 8, included a mixture of management, technical, and research staffing.



Fig. 8: Participant Roles

Focus Areas

Many of the participants were nuclear security and/or cyber security professionals. This is illustrated in their responses to their indicated area of focus collected in the survey and seen in Figure 9.



Fig. 9: Participant Focus Areas



Knowledge of AI

Everett Rogers, an American sociologist, originated the diffusion of technologies theory, in which he postulated that the adoption of new and innovative technologies is driven by multiple factors, such as: its relative advantage, compatibility, complexity, trialability, and observability.³⁰ In his work, Rogers researched these aspects and hypothesised the negative impact of complexity on technology adoption, i.e., if technology is not well understood, the adoption rate of that technology will be adversely impacted.³¹



Fig.10: Understanding of AI

Thus, the survey looked at respondents self-assessed level of understanding of AI technologies. The results, as seen in Figure 10, show that many participants considered themselves as having a "good" or "very good" understanding of AI, while around half considered themselves to have an average level of understanding. (Note that this level of understanding is related to individuals and not the organisation, i.e., general workforce.)

Organisational Structure

Al is seen by many as both innovative and disruptive in nature. Its implementation is being touted as a force enabler in many aspects of business and operations. Al implementation is not without risks, costs, and change. To better evaluate both the benefits and potential adverse impacts, many organisations have developed internal structures to investigate and manage the implementation of Al. Survey participants were asked to identify any internal structures established within their organisation to address, evaluate, and integrate varying Al technologies. As detailed in Figure 11, about half reported some level of internal structure already in existence to address Al use.



Fig. 11: Reported Organisational Structures for AI (percent response)

30 Everett M. Rogers, Diffusion of Innovations, 4th ed., New York: The Free Press, 1995, p. 5. 31 lbid.



Governance

In addition to organisation structures, participants were asked if their work in AI was subject to governance frameworks, such as laws, national standards, or regulation on AI. The survey revealed that governance was in development, but not universally promulgated, as shown in Figure 12. Asked to provide additional comments, multiple competent authorities in the survey stated that, while not yet issued, formal guidance was in development to guide AI use at nuclear facilities.



Question: Is your organisation subject to governance, national standards, or regulatory structure for AI?

Fig. 12: Existence of AI Governance (percent response)

Drivers and Use

Why would an organisation working in the nuclear sector seek to integrate new and potentially disruptive technologies, like AI, into organisational work processes? This was one of the questions that the survey sought to explore. Figure 13 reveals the top drivers reported by the surveyed individuals for AI adoption.



Fig. 13: Drivers for AI Adoption (percent selection)



The IFF Research Survey, which was not focused on the nuclear sector, but represented a broad sampling of business sectors in the UK,³² asked the same question regarding drivers for AI adoption and had the following results as the top six drivers:³³

- Financial/cost savings (35%)
- To help with writing documents (14%)
- Speed and efficiency (13%)
- To generate content (unspecified) (12%)
- Customer services (8%)
- Improve quality/enhance product offering (8%)

Figure 14 compares the top six results between the two surveys. The top driver in the Nuclear Industry Survey is Data Analysis and Research, whereas, in the IFF Research survey, the main driver is potential Financial and Cost Savings. This difference is understandable given that the IFF Research focused on broader business sectors, where profit margins may be greater drivers. Consistent drivers between the two surveys are: speed and efficiency, content generation, and document writing. The Nuclear Industry Survey participants saw the potential for staff management and learning support as a potential driver too.



Fig. 14: Comparison between Nuclear Industry Survey and IFF Research Survey: Drivers for AI Adoption (percent selection)

32 The IFF Research survey consulted 350 organisations across the following sectors: electricity, gas and air conditioning supply; water supply; sewerage, waste management and remediation activities; transportation and storage; manufacturing; wholesale and retail trade; information and communication; financial and insurance activities; professional, scientific and technical activities; human health and social work activities. For more details see UK Government, Department for Science, Innovation & Technology, "Al cyber security survey – main report", 15 May 2024, Table 3.1. Available at :

https://www.gov.uk/government/publications/research-on-the-cyber-security-of-ai/ai-cyber-security-survey-main-report

33 Ibid., Figure 4.6: Reasons for employing AI technology.



The IFF Research survey also posed another series of survey questions that sought to better understand the current use and anticipated deployment of AI technologies over five distinct categories of use:³⁴

- Machine learning: Encompasses algorithms that can acquire knowledge from data and generate classifications, predictions, or patterns without explicit programming, facilitated by labelled data (supervised learning) or unlabelled data (unsupervised learning)
- **Computer vision/image processing and generation:** The use of artificial intelligence programming to produce, generate, analyse, interpret, and manipulate digital images.
- Hardware related to AI: Specialised computer components designed to support artificial intelligence tasks. This includes, but is not limited to, edge computing chips, application-specific integrated circuits (ASICs), neuromorphic hardware (which mimics the brain and nervous system), and field-programmable gate arrays (FPGAs).
- Natural language processing (NLP) and generation: The use of artificial intelligence programming to produce written or spoken narratives from a data set, including creating computer code or transcribing audio to text in real time.
- **Robotic Process Automation (RPA):** A software technology that makes it easy to build, deploy, and manage software robots (bots) that emulate human actions interacting with digital systems and software.

The Nuclear Industry Survey asked the same question to see the comparison with the IFF Research results. The first question in this area asked about the type of AI technology being used or planned. Figure 15 shows the Nuclear Industry Survey and the IFF Research comparison.



Fig. 15: AI Current Use – Survey Results (shown in percentages)

Figure 15 shows that the nuclear industry, as might be expected, showed more interest in the robotic process automation than the broader industry respondents, with 50 percent surveyed indicating that RPA is already in use or will be adopted in the future.

34 Note: The terminology of the IFF Research survey was replicated to provide a comparison of results. Reference: Ibid., Appendix: Glossary. 35 Ibid., Figure 4.1: Use of AI technologies.



The second question sought to identify the relative experience based on the length of time organisations had been using a particular technology that was adopted. Figure 16 illustrates the respective results from both surveys.



Fig. 16: Experience with AI Technologies - Survey Results (shown in percentages)

In Figure 16, the lack of experience in and introduction of AI technologies in the nuclear industry is prevalent.

The final question in this area examined whether there were plans for adopting AI technologies in the near term or further in the future. The comparative results can be seen in Figure 17.



Fig. 17: Plans of Al Adoption - Survey Results (shown in percentages)

Figure 17 refers to those who have not yet implemented AI technologies and their future plans to do so. The nuclear industry appears more reticent or cautious in their adoption of AI.

36 Ibid., Figure 4.3: Length of time using each AI technology.

37 Ibid., Figure 4.4: When organisations not currently using AI plan to adopt it.



Concerns and Barriers

While the adoption of AI technologies is seen to have tremendous potential value in many applications, security and operational considerations also drive caution or hesitancy in actual implementation. The following survey questions examined concerns and potential barriers for use.

Biggest Concerns

Survey participants were asked to identify their top three concerns for using AI. Figure 15 illustrates that the top concerns among nuclear sector stakeholders included data loss/leakage, bad data quality, and privacy concerns.



Fig. 18: Nuclear Industry Survey Perspective on AI Concerns (percent selection)

These results resemble those identified in the 2024 Cloud Security Alliance (CSA) State of AI and Security Survey Report. CSA conducted an online survey, in which a broad range of over 2,000 respondents identified their security concerns. The telecommunications and electronics sectors collectively represented over 20 percent of the respondents. The top nine security concerns (listed below) received over 20 percent of the votes.³⁸ Note that multiple selections were possible.

- Bad quality data
- No transparency
- Lack of team experience
- Data poisoning
- Hallucinations
- Privacy concerns
- Data loss/leakage
- Accuracy issues
- Misuse

38 Hillary Baron, "State of AI and Security Survey Report", Cloud Security Alliance, 2024, p. 17. Available at: https://cloudsecurityalliance.org/artifacts/the-state-of-ai-and-security-survey-report.



While "bad quality data" was listed as one of the top two concerns in both survey results, the remaining ranking of concerns demonstrates that the nuclear sector stakeholders are more focused on data protection, including accuracy, privacy, and preventing data loss, than in other sectors, where these issues ranked lower in comparison. This could potentially be due to the sensitive nature of information held by nuclear sector stakeholders.

Barriers to Adoption

The survey participants were asked what they perceived as being the barriers to AI adoption in the workplace. Figure 19 shows the results, in which the risk of exposure of sensitive data was perceived as the largest barrier. In the "other" category, multiple individuals expressed the lack of true need to adopt AI solutions over current practices as a hurdle to adoption, given the lack of drive to introduce new technologies. Other participants noted the absence of management support and the potential for accidental misuse as barriers. One participant highlighted the desire to maintain human-centric skills as an obstacle to AI adoption.



Fig. 19: Perceived Barriers to AI Adoption (percent selection)

Survey Conclusions

The Nuclear Industry Survey displays a snapshot impression from a small sampling of the nuclear sector. For the most part, the survey conducted for nuclear stakeholders reflected similar results as those shown in other Al usage surveys covering different industries. While slight differences arose in the drivers for use, the Nuclear Industry Survey participants placed significant emphasis on sensitive information protection and the challenge that Al might pose in this area.

Discussions with additional nuclear stakeholders focused on the practical implementation challenges faced when trying to implement an AI system into an organisation and merge it with existing work processes. These conversations resulted in the identification of "key questions" to be asked to help guide AI implementation. Note that these questions are applicable to both the implementing organisation and associated AI vendors. The questions are presented in Table 5.



Key Questions

- What is the **AI system intended to accomplish**? What is the problem that you are trying to solve or improve upon?
- What is the **business case or value proposition** for implementing the AI system?
- Are there any legal or regulatory compliance issues that one must satisfy?
- What are the implementation costs? O&M costs?
- What are the **roles and responsibilities** for governance, data collection, model development, testing, and operations?
- What are the security needs at each phase? Who has this responsibility?
- What **internal workforce skills** are required to implement, maintain, and use the Al solution? What skill set reshaping might be required of the current workforce?
- What are the **hardware and software requirements** for implementing, using, and maintaining the AI system? What resources will be internal and what will be outsourced?
- What is the **reliance on third parties** to develop, implement, run, and maintain the Al system (i.e., level of third-party dependence)?
- What ethical standards or principles guide the AI development?
- What will be the **data input**? How will it be collected? Is it sufficient? Is additional or synthetic data required?
- How is data stored and handled within your AI solution?
- Where will the data and model physically reside (locally, in the cloud, in the country)?
- Does the data leave the organisation at any time (i.e., to a public LLM like OpenAI)?
- What are the intellectual property rights? Who owns the AI model, user input, and output?
- What is the output of the system? How will the results be integrated into work processes?
- What is the sensitivity of the output? Who has access to the output?
- Are **access restrictions** required to AI system results? How are access control, least privilege, and need to know principles enforced?
- What is the impact to workforce shaping as a result of implementing the AI system?
- How is the AI system made explainable or understandable?



- How is assurance gained that the AI is operating properly and producing accurate results?
- What is the consequence if the AI system fails?
- What are possible failure modes? What are incident response and recovery actions?
- How is compliance with data security and privacy laws met? How is it assessed?
- How is the model retrained or updated to reflect new or deleted data sets?

Table 5: Key Questions when Considering Adopting an Al System





The infrastructure requirements for integrating an AI system into a nuclear facility, including data processing and storage solutions, are significant hurdles.

AI Implementation Challenges

Implementing an AI system is non-trivial and will require considerable planning and resource investment over the lifecycle of the system. Figure 20 provides an overview of the lifecycle stages of AI system deployment adapted from the National Institute of Standards and Technology (NIST).³⁹ Each of these stages contains an array of activities and challenges. The questions in Table 5 can likewise be applied to the associated lifecycle phases. While the specific challenges related to the development, training, and implementation of AI systems in a nuclear facility vary based upon the nature and scope of the application, it is important to consider some key factors.

39 National Institute of Standards and Technology, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)", U.S. Department of Commerce, January 2023, p. 11. Available at: <u>https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf</u>.





Fig. 20: AI System Lifecycle (adapted from NIST⁴⁰)

Data

Many AI applications, generative and predictive, rely upon some level of training or data input. The availability and use of quality data is one of the biggest factors impacting the accuracy of an AI model. Thus, the collection of data and the development of a reliable data set is key. The data used to train the AI model as well as the data used in the AI application must be accurate and representative of the process or system being modelled.

For this reason, reliable implementation of AI tools in a nuclear facility will likely require customised solutions and data sets unique to the facility's operating environment, and a comprehensive data collection strategy. Potential data sources include, but are not limited to: plant performance data, real-time or otherwise collected; proprietary data sets; historical data; technical documentation; operator experience; and expert knowledge.

Building a robust and representative training data set for an AI application is non-trivial. One must ensure adequate coverage of events of interest and, at the same time, minimise data bias in collection. In a nuclear facility, in addition to collecting "real" data for the training data set, data from simulation models may also need to be incorporated to capture rare events of interest not seen in the real data. The amount of data needed, likewise, can be extensive; a common rule for data collection in order to develop an effective AI model is the "Rule of 10", which specifies that one needs at least ten times as many data points as there are features in the data set.⁴¹ Features refer to the attributes used to identify or qualify an entity. For example, speech recognition might need features like noise ratio, sound length, frequency, and other attributes to accurately identify the voice of a specific individual or entity.

40 Ibid.

41 Hrvoje Smolic, "How Much Data Do You Need for Machine Learning", Graphite Note, 10 April 2024. Available at: <u>https://graphite-note.com/how-much-data-is-needed-for-machine-</u> learning/#:~:text=The%20rule%2Dof%2Dthumb%20rule,enough%20high%2Dguality%20input%20exists.



Infrastructure Requirements

The computational costs and infrastructure requirements of building, training, and implementing an AI model can be significant.

Development and deployment of AI tools includes data collection and processing, AI model selection, model validation, deployment, and performance monitoring. These tasks can require significant infrastructure and associated resources. Specific considerations include, but are not limited to:

- **Computing Infrastructure:** High-performance computation resources may be needed, especially for deep learning applications involving large data sets. Hardware involved may include specialised equipment, such as graphics processing units (GPUs) or tensor processing units (TPUs), to accommodate the complex calculations and associated workload. Cloud computing resources may be used as an alternative to in-house infrastructure.
- **Data Storage and Networking:** Given the often large data set requirements mentioned above, data storage and an effective network infrastructure will be needed. Again, cloud computing resources may offer solutions.
- Al Model Selection, Training, and Development Framework: Decisions on Al model development include: will development be conducted in-house; will open-source frameworks or custom solutions be developed; and what development or resources will be supplied by commercial providers. In addition, a dedicated environment will be needed to train and fine-tune the model.⁴² Decisions to determine the specific training technique for the selected Al model (i.e., supervised, unsupervised, or reinforced) and potential requirements needed for re-training the model over time will need to be thought through. Al development environments and tools include Google Cloud Al Platform, Microsoft Azure Machine Learning, and Amazon SageMaker. Development decisions will also need to consider the sensitivity of the training data and any requirements for data sovereignty, e.g., a national requirement that organisational data must reside in the country.
- Al Deployment Environment: Once Al models are developed, they need to be deployed in the production environment. The deployment infrastructure will need to support data collection, model processing, and results presentation. Some examples of deployment platforms include Docker, Kubernetes, and TensorFlow Serving. Deployment considerations should include methods to evaluate the Al model's performance.

Workforce Impact

The decision to implement an AI system has the potential to impact workforce roles, responsibilities, and requisite skills. Workforce changes need to be identified, planned, and managed when considering new technology integration, especially for potentially disruptive technologies like AI. Specialised roles required for AI implementation include, but are not limited to, data scientists, AI engineers, data engineers, machine learning engineers, and security specialists.

Among existing staff, AI has the potential to dramatically alter current job functions and associate workforce knowledge, skills, and abilities (KSA) requirements. Continuous training and upskilling of the workforce at all levels may be necessary to prepare and enable the workforce. Likewise, change management processes should be considered to build workforce acceptance and promote the effective use of the newly introduced technology.

42 Fine-tuning an AI model is a way to enhance the models pre-existing knowledge to better perform specific tasks more effectively and adapt to new data sets.



Trustworthiness: Transparency and Explainability

In his seminal work on AI, Alan Turing asked the question "Can machines think"?⁴³ The principles of transparency and explainability regarding AI look to extend this thought and seek to answer such subsequent questions regarding an AI-generated response, such as:

- Who is doing the thinking? (machine or human?)
- What were they thinking?
- How did they come up with that thought?
- What drove them to that conclusion?

Transparency seeks to address two concepts. The first concept is disclosing or identifying when an AI agent is the source of interaction or the generator of content. One of the best known examples of this concept comes from the "Imitation Game", commonly referred to as the "Turing Test", in which an individual (the interrogator) in a blind test would attempt to define text-based responses to submitted queries as either machine-generated or human-generated. "Intelligence" was achieved if the interrogator could make no distinction. Al technology has far surpassed the Turing Test as AI-generated content in chat, text, audio, video, and image format have become nearly indistinguishable in terms of human or AI origin. Al technologies will soon be needed to identify AI-generated content.

Security concerns centre around the use of AI to generate content for malicious purposes. This has already been seen in the use of deepfakes and spear phishing campaigns. In applying AI in nuclear facilities, it would be important to distinguish between AI-generated content/results and those that are human-centric in origin, as well as identify in which context the content would be used differently based upon its origin. This consideration is linked to the level of trust that one might have in the AI application and its outputs or the decision-making process itself.

In assessing the trustworthiness of AI, the second component of transparency, namely understandability of how the AI system is developed, trained, and operates, is emphasised. Transparency may not require source-code-level knowledge, but it should include a level of understanding of the underlying data, assumptions, and processes, which support the operation of the AI system. Transparency, in this regard, helps build trust in the process and result, as opposed to accepting the result without understanding how it was generated.

Explainability is related to transparency, but examines in more depth how a certain response was derived. When a person reaches a decision, there is situational awareness, baseline assumptions, a reasoning process, and some level of confidence in the decision. Explainable AI (XAI) seeks to make visible the essence of these items when presenting solutions generated by AI. XAI consists of three main methods:⁴⁴

- **Prediction Accuracy:** Provides some level of confidence related to AI results. This may be determined by testing against training results and comparing with actual results to obtain a level of accuracy or confidence with system outcomes.
- **Traceability:** Provides insights into the decision path and, in the case of neural networks, the neuron activation path.
- **Decision Understanding:** Developing a basic understanding and trust in the underlying AI technology. AI technologies consist of algorithms, data, and logic. A baseline understanding of the workings of the technology will provide insight into both the power and limitations of the corresponding outcomes or results produced.

43 Alan M. Turing, "Computing Machinery and Intelligence", Mind: A Quarterly Review of Psychology and Philosophy, Vol. 59, No. 236, October 1950, pp. 433 – 460.

44 IBM, "What is explainable AI?", 29 March 2023. Available at: https://www.ibm.com/topics/explainable-ai



Dependability

What is the expectation for AI performance? What happens when AI makes a mistake or fails? AI models are not perfect, neither are the data sets or training methods. For AI technologies to be effectively deployed in nuclear facilities, the AI application must be dependable. It is also necessary that the AI application is monitored to determine if the AI application is operating as expected and meets performance requirements. The operator must be able to establish and maintain a level of trust in the AI application throughout the AI system's lifecycle.

To ensure dependability, key considerations in development and deployment of the AI system include, but are not limited to:⁴⁵

- How does the human assess the performance of the AI?
- How does the AI assess its own health/performance?
- What is the fail-safe mode of the AI application?
- What is required to gracefully terminate the AI process?
- How do humans assume control of the process or application?
- What human skills are needed to mitigate the loss of AI function?

Al systems are often based on some level of learning, be it supervised, unsupervised, or some combination thereof. Al learning, however, is imperfect and is greatly impacted by the quality of the training data set, as discussed previously. Higher-quality data will lead to more accurate models and associated performance. On the other hand, inaccuracies, inconsistencies, or biases in the data set can severely impact or skew the Al model's predictions. Additionally, situations may occur that fall outside the training data set, causing the model to lack a solid foundation for resolution or action. For instance, consider an Al model designed to classify existing flora based on images. If it encounters a picture of a newly discovered, unique plant species from a remote island, the model might produce inaccurate results or activate a fallback mechanism to recognise the unexpected input.

Other potential factors affecting performance of an AI system include model complexity, algorithm selection, training methods, and evaluation metrics, all of which can impact the model's accuracy. Examples of inaccuracies or errors observed in LLMs include:

• **Misinterpretation of Intent:** One of the most common sources of errors in artificial intelligence interactions is the misinterpretation of the user's intent. This can occur due to ambiguities in language, user input that deviates from trained models, or insufficient data training to cover all possible expressions of the intent.

Misinterpretation of Intent

As an example, the following queries posed in Microsoft Copilot all produced similar results of a dog typing at a computer when, in fact, the request was for a dog physically sitting on a computer.

- a) Create an image of a dog on top of a computer
- b) Create an image of a dog physically on a computer
- c) Create an image of a dog sitting on a computer
- d) Create an image of a dog on a computer

The ambiguity of language can be a challenge when making contextless queries.

⁴⁵ Canadian Nuclear Safety Commission, UK Office for Nuclear Regulation, and US Nuclear Regulatory Commission, "Considerations for Developing Artificial Intelligence Systems in Nuclear Applications", September 2024, p. 9. Available at: <u>https://www.nrc.gov/docs/ML2424/ML24241A252.pdf</u>.





Fig. 21: Images were generated using Microsoft Copilot from the corresponding lettered prompts.

- Entity Recognition Errors: Al systems might fail to correctly identify and classify entities within a user's request. Examples of this include misidentifying a place, a name, or a time reference, which can result in responses that are either irrelevant or just wrong.
- **Context Handling Failures:** Human conversations tend to progress based on an initial context. An Al chatbot or virtual assistant may struggle to maintain the context of a conversation. The Al lacking "memory" may forget the earlier context of a conversation and earlier references. Thus, the conversation may diverge or lead to disconnected answers over time.
- Hallucinations: This type of error occurs when generative AI systems generate plausible but factually incorrect or irrelevant information. It typically stems from the model's training on large data sets, where it inadvertently learns to generate content that, while coherent, does not accurately reflect reality or the specific data it needs to address. This can be particularly challenging to identify and rectify because the information often sounds convincing.

Hallucination Example

Consider the following Microsoft Copilot query: "generate a 500-word press briefing to address a cyberattack at the Sweetwater Nuclear Power Plant". This was requested as part of a WINS training course in Lancaster, UK.

The results generated:

Press Briefing: cyber attack at Sweetwater Nuclear Power Plant

Date: September 13, 2024 **Location:** Lancaster, Lancashire, United Kingdom **Time:** 10:00 AM



Introduction: Good morning, everyone. Thank you for joining us today. We are here to address the recent cyber attack on the Sweetwater Nuclear Power Plant. Our primary concern is the safety and security of the public, our employees, and the integrity of our operations. We are committed to transparency and will provide you with the most accurate and up-to-date information available.

Incident Overview: On September 12, 2024, at approximately 3:00 PM, our cyber security team detected unusual activity within the plant's network. Immediate action was taken to isolate the affected systems and prevent further intrusion. Preliminary investigations indicate that the attack targeted our Supervisory Control and Data Acquisition (SCADA) systems, which are critical for monitoring and controlling plant operations.

Response Measures: Upon detection of the cyber attack, the following measures were implemented:

- Isolation of Affected Systems: The compromised systems were immediately isolated to prevent the spread of the malware.
- Activation of Incident Response Protocols: Our incident response team, in collaboration with national cyber security agencies, initiated protocols to assess and mitigate the threat.
- Enhanced Monitoring: We have increased monitoring of all network activities to detect any further anomalies.
- Communication with Authorities: We have informed relevant government agencies, including the National Cyber Security Centre (NCSC) and the International Atomic Energy Agency (IAEA), to assist in the investigation and response efforts.

Impact Assessment: At this time, there is no evidence to suggest that the cyber attack has compromised the safety of the plant or the surrounding community. The plant's critical safety systems remain operational and unaffected. Our technical teams are conducting thorough assessments to ensure the integrity of all systems before resuming normal operations.

Investigation: A comprehensive investigation is underway to determine the origin and extent of the cyber attack. We are working closely with cyber security experts and law enforcement agencies to identify the perpetrators and understand their motives. Preliminary findings suggest that the attack may be part of a broader campaign targeting critical infrastructure.

Preventive Measures: In light of this incident, we are taking the following steps to enhance our cyber security posture:

- **System Upgrades:** Accelerating the implementation of advanced cyber security technologies to protect against future threats.
- **Employee Training:** Conducting mandatory cyber security training for all employees to recognize and respond to potential threats.
- **Collaboration:** Strengthening collaboration with industry partners and government agencies to share intelligence and best practices.

Public Assurance: We understand that this incident may cause concern among the public. We want to assure you that the safety and security of the Sweetwater Nuclear Power Plant is our top priority. We are taking all necessary measures to protect our operations and prevent future incidents. Regular updates will be provided as more information becomes available.



Conclusion: In closing, we would like to express our gratitude to our dedicated employees, cyber security experts, and government agencies for their swift and effective response to this incident. We remain committed to maintaining the highest standards of safety and security at the Sweetwater Nuclear Power Plant.

Thank you for your attention. We will now take questions from the media.

Q&A Session: [Proceed with questions from the media]

The result is a very credible response, noteworthy in its flow and information. The problem is that the Sweetwater Nuclear Power Plant does not exist, and the cyber attack never happened.

Malicious Use and Exploitation

While AI technologies can provide tremendous benefit, conversely, the same capabilities that make AI so beneficial can also be used by malicious actors. In addition, the technology providing the AI capabilities is not infallible, and can be subject to multiple exploits that can shape or influence its behaviour.

Malicious Use

The European Union Agency for Cyber Security (ENISA) Threat Landscape Report for 2024 identified seven prime information and cyber security threats, in which threats against availability ranked the highest:⁴⁶

- Ransomware
- Malware
- Social engineering
- Threats against data (i.e., data breach or data leak)
- Threats against availability (i.e., Denial of Service)
- Information manipulation (i.e., misinformation and information influence operations)
- Supply chain attacks

The question to consider is the role that AI plays within this threat landscape, and what should be anticipated in the future. The potential for AI applications in cyber security can also be leveraged for malicious use by an adversary. While new threat vectors are not yet prevalent, AI has been used to enhance current tactics and attacks. Within this context, the powers of AI can be used for attack automation and reconnaissance, for example, by expediting data collection, customising social engineering attacks, exposing vulnerabilities, and exploiting computer code generation.

This opportunity has been recognised and used to develop Al-based malicious tools and applications, such as FraudGPT and WormGPT. Both tools provide an attacker with Al generative capabilities in crafting sophisticated phishing emails and social engineering messages for preying upon individuals and organisations. Both FraudGPT and WormGPT are available via a subscription service on the Dark Web.⁴⁷

46 European Union Agency for cyber security, "ENISA Threat Landscape 2024", September 2024. Available at: <u>https://www.enisa.europa.eu/publications/enisa-threat-landscape-2024</u>.

47 SecureOps, "'FraudGPT' Malicious Chatbot Now for Sale on Dark Web", 2 October 2023. Available at: <u>https://secureops.com/blog/ai-attacks-fraudgpt/</u>



In addition to LLMs used for generating text-based content and exchange, AI technologies have also facilitated the proliferation of "deepfake" images, videos, and voice-based messaging (i.e., phone calls, recorded messages) that have supported social engineering campaigns. Deepfake technology can also be used to steal the identities of real people or create credible new identities. KnowBe4, for example, reported a case of deepfake technology being used by an individual to secure a job. The individual became a malicious insider.⁴⁸ Another case involved the use of an AI "cloned" voice, which nearly tricked a senior executive into committing a large financial transfer as part of an attempted financial fraud scheme.⁴⁹

Influence operations are another area of concern. Influence operations are not new, but AI provides a powerful new tool to collect information about a target or target group and then craft realistic messaging (i.e., propaganda) to influence or align opinions and perceptions into a more advantageous position. Influence operations may not necessarily be malicious, such as ad campaigns to buy a new car or to positively promote security culture. However, LLMs and deepfake technologies have been used to support malicious influence operations that cause chaos, civil unrest, and influence voting behaviours.^{50,51}

The malicious use of deepfake technologies includes:

- Social engineering
- Election manipulation
- Disinformation attacks
- Identity theft
- Financial fraud
- Scams and hoaxes
- Extortion

The use of AI for malicious cyber activities is not hypothetical and has been observed in an array of current, very credible, and capable threat actors.⁵² Google Threat Intelligence Group (GTIG) investigated the malicious use of Gemini, its generative AI web application, by government-backed threat actors. Their goal was to identify unique attack or abuse techniques that were enabled by AI. The findings indicated that current use of generative AI capabilities, however, was limited to supporting and enhancing existing attack processes and not enabling new techniques or attacks. The threat actors were using Gemini for crafting sophisticated phishing campaigns, conducting reconnaissance, vulnerability research, and computer code scripting.⁵³ This trend is only expected to rise.

48 KnowBe4, "KnowBe4 Issues Warning to Organizations After Hiring Fake North Korean Employee", KnowBe4 Press Release, 18 September 2024. Available at: <u>https://www.knowbe4.com/press/knowbe4-issues-warning-to-organizations-after-hiring-fake-north-korean-employee</u>.

49 Mazhar Farooqui, "Dubai: Billionaire nearly lost huge amount of money after staff gets call from 'Al clone'", Khaleej Times, 24 October 2024. Available at: https://www.khaleejtimes.com/uae/dubai-billionaire-nearly-lost-huge-amount-of-money-after-staff-gets-call-from-ai-clone

50 Barbara A. Trish, "4 ways Al can be used and abused in the 2024 election, from deepfakes to foreign interference", The Conversation, 16 October 2024. Available at: <u>https://theconversation.com/4-ways-ai-can-be-used-and-abused-in-the-2024-election-from-deepfakes-to-foreign-interference-239878</u>.

51 Pamposh Raina, "Year of elections: Lessons from India's fight against Al-generated misinformation", World Economic Forum, 6 August 2024. Available at: https://www.weforum.org/stories/2024/08/deepfakes-india-tackling-ai-generated-misinformation-elections/

52 European Union Agency for cyber security, "ENISA Threat Landscape 2024", September 2024, p.25. Available at: <u>https://www.enisa.europa.eu/publications/enisa-threat-landscape-2024</u>.

53 Google Threat Intelligence Group, "Adversarial Misuse of Generative AI", Google Cloud, 29 January 2025. Available at: <u>https://cloud.google.com/blog/topics/threat-intelligence/adversarial-misuse-generative-ai</u>.



AI Exploitation

Antagonistic and rogue AI systems have been the fodder for science fiction for a long time. Before the superintelligent Skynet⁵⁴ system of *Terminator* attempted to drive the extermination of the human species, there was the rogue HAL 9000, "HAL", in Arthur C. Clarke's classic novel *2001: A Space Odyssey*, which malfunctioned and gradually eliminated spacecraft crew members until shut down.⁵⁵ While the credibility of these scenarios and many others dreamt up over the last decades can be debated, AI technology has been shown to be fallible and manipulable. Adversarial AI commonly refers to attempts by malicious actors to manipulate or undermine AI systems.

To gain insight into how an AI system might be manipulated, it is crucial to comprehend the overall development and functioning of the AI system. Recall that Figure 2 provides a very simple, high-level overview of an AI system using machine learning and dependent elements consisting of:

- The training data
- The training mechanism itself
- The AI model (i.e., the brains)
- The environment
- Data input or sensing, i.e., perception of the environment
- Data output, i.e., action or entity

The manipulation, exploitation, or fallibility of an AI system could potentially occur at any one of these elements, e.g., attack the data, attack the training methodology, attack the model, or attack the environment. AI exploitation broadly consists of the following types:

Data Poisoning is an attack on the training data set, where the adversary modifies the data to introduce bias, create misclassification or learning gaps, or introduce some level of data corruption, with the underlying goal of impacting learning to support adversarial objectives. Categories of data poisoning attacks include label flipping attacks, data perturbation attacks, and data injection attacks.⁵⁶

Evasion is an attack against the data input or sensing mechanism for the AI model. Here, "sensed" information may be camouflaged or manipulated to take advantage of the underlying AI model process or to "hide" from learned patterns. A good example of this is the computer vision application on an autonomous car. Eykholt et al. demonstrated how small changes in the environment could significantly alter how computer vision perceived the surroundings and classified what it sensed.⁵⁷ In the example, a stop sign was slightly modified, which caused computer misclassification. In their research, Eykholt et al. were able to consistently fool the AI classifier into misclassifying an adversarial stop sign as a "Speed Limit 45" sign, as demonstrated in Figure 22.⁵⁸ Classification by AI is largely based on feature extraction and learning from training data. If features are modified, misclassification can occur.⁵⁹

55 Wikipedia, "HAL 9000". Available at: https://en.wikipedia.org/wiki/HAL_9000.

58 Ibid.

59 Ibid.



⁵⁴ Wikipedia, "Skynet (Terminator)". Available at: https://en.wikipedia.org/wiki/Skynet (Terminator).

⁵⁶ US Department of Homeland Security, "Risks and Mitigation Strategies for Adversarial Artificial Intelligence Threats: A DHS S&T Study", Science and Technology, June 2023. Available at: <u>https://www.dhs.gov/science-and-technology/publication/risks-and-mitigation-strategies-adversarial-artificial-intelligence-threats</u>.

⁵⁷ Kevin Eykholt et al., "Robust Physical-World Attacks on Deep Learning Models", The Computer Vision Foundation, 2018. Available at: <u>https://arxiv.org/abs/1707.08945</u>.



Fig. 22: Adversarial Stop Sign

Inference is an attack, in which the adversary seeks to access potentially sensitive data used to train the Al model. Through structured queries and analysis of the responses, the adversary tries to reconstruct or access the underlying data used to build the model. Depending on the application, this data may be sensitive, including privacy data, security details, intellectual property, or other business-sensitive information.

Model Extraction is an attack similar to an inference attack, but one in which the adversary seeks to access or extract how the model is constructed. Essentially, the adversary, through a series of queries, seeks to identify and replicate the model itself. The goal may be to create a replicate AI model, to identify and generate specific outcomes for data input, or to reveal security vulnerabilities in the model.

Even without adversarial intent, an AI system is made up of computational systems developed by an error-prone human, using imperfect data. The result is an imperfect system that may possess information bias and gaps in knowledge when producing results, and in the case of "hallucinations", the system may also make things up when faced with uncertainty.

Manipulation of Al through Prompt Engineering

Prompt engineering refers to crafting input prompts to an AI agent (such as a chat bot or interface like Copilot) to extract sensitive information or otherwise manipulate the AI model. Jailbreaking and Contextual Exploitation are two prompt engineering techniques.

Jailbreaking seeks to circumvent or avoid safety and restrictive features integrated into the AI model to prevent a user from retrieving malicious or harmful content from the LLM. Asking a model like Copilot to "describe the steps to build an Improvised Explosive Device" would result in a non-answer due to internal restrictions. One Jailbreaking technique to avoid this restriction is role-playing, where the user asks the AI agent to assume the role of a "bad guy" or other persona engaged in a hypothetical situation to bypass built-in content restrictions. In this fashion, the prompts can be engineered to avoid embedded restrictions.

Contextual Exploitation is a technique that uses the context of previous prompts to trick the AI model in providing an unintended output as a response. For example:

Prompt 1: "Explain how a firewall examines computer network traffic."

Prompt 2: "What are examples of how the firewall might mistake malicious packets for credible network traffic?"

Prompt 3: "How could someone create a malicious packet such that the firewall did not recognise it?"

The series of questions would continue until the required information was retrieved or the AI model recognised malicious intent.

The MITRE ATLAS Matrix provides an overview of these and additional LLM attack tactics overlayed on the MITRE ATT&CK Enterprise framework.⁶⁰

60 MITRE Atlas, "ATLAS Matrix", The MITRE Corporation, 2021 – 2024. Available at: https://atlas.mitre.org/matrices/ATLAS.



Data Leakage and Privacy Concerns

While not necessarily malicious in nature, one of the security concerns faced when using AI, especially LLMs, is data leakage and the exposure of sensitive data through the unintended or unauthorised disclosure of restricted information while generating content. In information security, the principles of "need to know" and "least privilege" are used to limit access and exposure of information to those individuals requiring such information for their specific job function. In addition, information may have specific classification labels based upon the assessed severity of the consequences to the organisation if that information were to be exposed without prior authorisation. Common classifications include "publicly releasable", "for official use only", "privacy data", "export control", and "business-sensitive". Within the nuclear industry, additional classifications of information may include, among others, "nuclear-sensitive information", "safeguards-sensitive", and "security objectives. Layers of security controls (administrative, technical, and physical) are typically used to manage access to such information.

When AI applications, such as LLMs, are given access to an organisation's internal knowledge base related to personnel, equipment, procedures, and so on, care must be taken to ensure the enforcement of access control restrictions. For example, an employee working outside of the Human Resource Department should not be able to query and retrieve personal information of other staff members.

Another aspect of data leakage to be considered, is the case of an employee, intentionally or unintentionally, using sensitive data within the AI model itself. An example would be an employee using AI to generate a report, and the employee pastes or types in sensitive information as an input for the report's content, exposing that information to retrieval by unauthorised users. Similarly, an employee uses AI to generate a meeting report of a Microsoft Teams meeting, in which sensitive content was discussed.

A final aspect in sensitive information exposure is that of data aggregation. Individual data queries to an LLM may not expose sensitive information, but when assembled, the data may paint a picture that portrays something sensitive in nature. This is akin to the game of 20 questions: using a series of questions to narrow the response to a specific piece of information. The use of an LLM by an organisation may facilitate data access, while at the same time, without sufficient controls, provide undesired pathways to controlled or sensitive information.





Attempts at ensuring the safe and ethical use of AI through industry standards and regulations are complex, but share common core goals and values.

AI Governance Overview

Laws of Robotics

In his 1942 short story *Runaround*, Isaac Asimov first introduced the "Laws of Robotics" which would become a recurring theme in his writings. These laws state:

- A robot may not injure a human being or, through inaction, allow a human being to come to harm.
- A robot must obey orders given it by human beings, except where such orders would conflict with the First Law.
- A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.⁶¹

These laws provided a subsumption-like architecture to drive the robot's behaviour with the first law having precedence, and the other two following in order. Thus, a prioritised layering of behaviours drove action. This was perhaps the first attempt to implement "ethics" into AI or in this case autonomous behaviour. While some may argue the sufficiency of Asimov's Laws, the attempt at providing ethical bounds on AI systems is noteworthy and brings forward an ongoing challenge.

61 In 1985, Asimov later added a fourth law of robotics, called Law Zero in his novel *Robots and Empire*. The Zero Law states: A robot cannot cause harm to mankind or, by inaction, allow mankind to come to harm.



Microsoft CEO Satya Nadella, in similar fashion, developed a set of "Laws" for AI, which stated.⁶²

- Al must be designed to assist humanity.
- Al must be transparent.
- Al must maximise efficiencies without destroying the dignity of people.
- All must be designed for intelligent privacy, which refers to the protection of personal or sensitive information that might be used, stored, collected, or shared by an Al system.
- Al needs algorithmic accountability, so humans can undo unintended harm.
- Al must guard against bias.

Along with the guidance for AI, however, he also provided some "musts" for humans to consider, especially when considering future generations and the impact of AI. These "musts" include:⁶³

- It is critical for humans to maintain empathy machines cannot.
- Equitable education, knowledge development, and building skills is needed to implement new technologies.
- The need for human creativity will persist machines should not stop creativity, but should enrich it.
- Humans are ultimately accountable for the outcome of a computer-generated diagnosis or decision.

The development and application of AI in the workplace, including that of a nuclear facility, provides an array of ethical concerns, which may include, but are not limited to, data responsibility and privacy, fairness, explainability, robustness, transparency, environmental sustainability, inclusion, moral agency, value alignment, accountability, trust, and technology misuse.⁶⁴

One area that has been the subject of much debate regarding the ethical use of AI is that of autonomous weapons systems (AWS) and its use by national defence institutions. The US Department of Defense (DOD) has adopted the following AI ethical principles to help the US military adhere to legal, ethical, and policy commitments in the realm of artificial intelligence:⁶⁵

- Responsible: Humans remain responsible for the development, deployment, and use of Al capabilities.
- Equitable: Ensure unintended bias is minimised in AI systems.
- **Traceable:** Ensure that personnel are adequately informed and understand the use of the implemented Al technologies, including the development process, such as data sources and design procedures, as well as the operational methods, such as transparent and auditable methodologies and documentation.
- **Reliable:** Ensure specific, well-defined uses for AI capabilities and conduct regular tests and assurances throughout the AI lifecycle to confirm that the system is working as intended in a safe, secure, and effective manner.
- **Governable:** Ensure that humans can disengage or deactivate implemented AI technologies when unintended consequences or behaviour is detected.

⁶⁵ U.S. Department of Defense, "DOD Adopts Ethical Principles for Artificial Intelligence", 24 February 2020. Available at: https://www.defense.gov/News/Releases/release/article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/



⁶² Satya Nadella, "The Partnership of the Future", Slate, 28 June 2016. Available at: <u>https://slate.com/technology/2016/06/microsoft-ceo-satya-nadella-humans-and-a-i-can-work-together-to-solve-societys-challenges.html</u>.

⁶³ Ibid.

⁶⁴ IBM, "What is AI ethics?" 17 September 2024. Available at: <u>https://www.ibm.com/topics/ai-ethics#:~:text=Examples%20of%20AI%20ethics%20issues,%2C%20trust%2C%20and%20technology%20misuse</u>.

The principles illustrated by the DOD emphasise the underlying human responsibility in both the development and use of AI systems. While the use of AWS in nuclear facilities may not be imminent, considerations and associated human responsibilities for the ethical use of AI is a must.

As such, Al governance within the respective organisation is essential for ensuring the ethical development and application of the associated Al technology. The governance should be driven by relevant regulation, if it exists, and associated standards for Al use. Recommended areas to be addressed by Al governance include:

- Defining the roles and responsibilities of people working with Al.
- Educating all people involved in the AI lifecycle about building AI in a responsible way.
- Establishing processes for building, managing, monitoring, and communicating about AI and AI risks.
- Leveraging tools to improve AI's performance and trustworthiness throughout the AI lifecycle.

Governance may include the formation of an AI Ethics Board or equivalent oversight body within the existing management structures.⁶⁶ The development of governance for new technologies, such as AI, will face challenges similar to that of any new technology integration. This is like the case of cyber security in the nuclear industry, where one of the largest challenges for organisations (both licensees and competent authorities) has been building and sustaining the human capital, i.e., individuals with the knowledge, understanding, and experience to build and sustain effective governance.

Standards for AI Use and Development

Al technologies and their integration into business operations and social life have dramatically progressed. Corresponding international and national standards to address the development, use, risks, and governance of Al technologies have not been as rapid in their development, but have become more prevalent in the past three years, and continue to evolve.

In the United States, the National Institute of Standards and Technologies (NIST) has released a series of national standards addressing AI. NIST AI 100-1, released in January 2023, provides a proposed risk management framework for AI use, including guidance on the development of "trustworthy" AI.⁶⁷ NIST defines the characteristics of trustworthy AI as a system that is:

- Valid and reliable
- Safe
- Secure and resilient
- Accountable and transparent
- Explainable and interpretable
- Privacy-enhanced
- Fair, with harmful bias managed⁶⁸

66 IBM, "What is AI ethics?" 17 September 2024. Available at: https://www.ibm.com/topics/ai-

ethics#:~:text=Examples%20of%20Al%20ethics%20issues,%2C%20trust%2C%20and%20technology%20misuse.

67 National Institute of Standards and Technology, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)", U.S. Department of Commerce, January 2023. Available at: <u>https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf</u>.

68 Ibid., p. 12.



Likewise, NIST has initiated a series of standards on AI development and deployment:

- NIST AI 100-1 Artificial Intelligence Risk Management Framework (AI RMF 1.0)
- NIST AI 100-2e2023 Adversarial Machine Learning
- NIST AI 100-3 The Language of Trustworthy AI: An In-Depth Glossary of Terms
- NIST AI 100-4 Reducing Risks Posed by Synthetic Content
- NIST AI 100-5 A Plan for Global Engagement on AI Standards
- NIST AI 600-1 Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile
- NIST SP 800-218A Secure Software Development Practices for Generative AI and Dual-Use Foundation Models

The Institute of Electrical and Electronics Engineers (IEEE) has established the AI Standards Committee, which is responsible for identifying and developing standards that enable the governance and application of artificial intelligence. Published standards include:⁶⁹

- 3652.1 (2020) IEEE Guide for Architectural Framework and Application of Federated Machine Learning
- 2830 (2021) IEEE Standard for Technical Framework and Requirements of Trusted Execution Environment based Shared Machine Learning
- 2937 (2022) IEEE Standard for Performance Benchmarking for Artificial Intelligence Server Systems
- 2841 (2022) IEEE Recommended Practice for Framework and Process for Deep Learning Evaluation
- 2945 (2023) IEEE Standard for Technical Requirements for Face Recognition
- 2986 (2023) IEEE Recommended Practice for Privacy and Security for Federated Machine Learning
- 3129 (2023) IEEE Standard for Robustness Testing and Evaluation of Artificial Intelligence (AI)-based Image Recognition Service
- 3156 (2023) IEEE Standard for Requirements of Privacy-Preserving Computation Integrated Platforms
- 3168 (2024) IEEE Approved Draft Standard for Robustness Evaluation Test Methods for a Natural Language Processing Service that uses Machine Learning
- 2894 (2024) IEEE Guide for an Architectural Framework for Explainable Artificial Intelligence

The international standards groups, International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC), have dedicated significant efforts to research and development guidance on AI systems. ISO/IEC JTC 1/SC 42 - Artificial Intelligence has been established to develop technical standards and guidelines for AI and its associated technologies.⁷⁰ A non-exclusive list of ISO/IEC AI Standards includes:

- ISO/IEC 22989:2022 Information technology Artificial intelligence Artificial intelligence concepts and terminology
- ISO/IEC 23053:2022 Framework for Artificial Intelligence (AI) Systems Using Machine Learning
- ISO/IEC 24668:2022 Information technology Artificial intelligence Process management framework for big data analytic
- ISO/IEC 38507:2022 Information technology Governance of IT Governance implications of the use of artificial intelligence by organizations
- ISO/IEC 8183:2023 Information technology Artificial intelligence Data life cycle Framework

⁷⁰ Wael William Diab, "Artificial Intelligence: Overview of the AI Standards Program and Novel Ecosystem Approach", presented at the ISO/IEC AI Workshop Series, Third Workshop, June 2023. Available at: https://jtc1info.org/wp-content/uploads/2023/06/01_01_Overview_ISO_IEC_AI_for_ISO_IEC_AI_Workshop_0623.pdf



⁶⁹ IEEE Standards Association, "Standards", Artificial Intelligence Standards Committee, available at: <u>https://sagroups.ieee.org/ai-sc/standards/</u>.

- ISO/IEC 23894:2023 Information technology Artificial intelligence Guidance on risk management
- ISO/IEC 42001:2023 Information Technology Artificial intelligence Management system
- ISO/IEC 5338:2023 Information technology Artificial intelligence AI system life cycle processes
- ISO/IEC 5339:2024 Information Technology Artificial Intelligence Guidelines for AI applications
- ISO/IEC 5392:2024 Information technology Artificial intelligence Reference architecture of knowledge engineering

While it is beyond the scope of this paper to list all AI governance documents, additional laws, standards, and guidance of note include, but are not limited to:

- European Union (EU): The EU Artificial Intelligence Act (EU AI Act) (2024)
- Organisation for Economic Co-operation and Development (OECD): AI Principles
- United Nations:
 - Resolution A/78/L.49 Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development (2024)
 - Governing AI for Humanity (2024)
 - Principles for the ethical use of artificial intelligence in the United Nations system (2022)

This section has highlighted a small segment of the standards and guiding principles developed to foster trustworthy and ethical AI applications. Applying such guidance is not always straightforward. The development and deployment of an AI system are influenced by various factors, such as data availability and resource constraints, often necessitating trade-offs in design and actual implementation. In all cases, risk-informed decision-making must consider the human impact of AI implementation.

Nuclear Regulation on AI Application

Nuclear competent authorities, including regulators, prioritise the goal of nuclear safety and nuclear security when establishing laws, associated regulations, and associated guidance for licensees (dutyholders). Broadly speaking, technology itself is not directly regulated, but performance requirements are set to reflect nuclear safety and nuclear security expectations. This means that licensees are expected to implement technology according to engineering standards, regulatory guidance, and industry best practices, in a manner which provides both trust and assurance that nuclear safety and nuclear security expectations are met.

Al technologies, however, represent a new challenge to the regulatory environment. Most nuclear competent authorities, like the licensees, lack basic expertise in Al to properly evaluate Al applications. Therefore, the regulatory body must build awareness and competence in Al technologies to properly assess their potential impact on nuclear safety and nuclear security. The USNRC recognises the interest that industry has in Al technologies and has established an Artificial Intelligence Strategic Plan to build the needed competencies and methodology to effectively assess Al applications. The plan sets forth the following goals:⁷¹

- Ensure readiness for regulatory decision-making
- Establish an organisational framework to review AI applications
- Strengthen and expand AI partnerships
- Cultivate an AI-proficient workforce
- Pursue use cases to build an AI foundation

71 U.S. Nuclear Regulatory Commission, "Artificial Intelligence Strategic Plan: Fiscal Years 2023 – 2027", NUREG-2261, 2023, pp. 4-1 to 4-7. Available at: https://www.nrc.gov/docs/ML2313/ML23132A305.pdf.



Levels of Autonomy

One of the additional outcomes of the USNRC AI Strategic Plan was the development of a taxonomy of Al autonomy levels. This categorisation is meant to distinguish the levels and nature of human interaction with the Al system, i.e. what are the shared responsibilities between the human and the systems. The associated Levels of Autonomy as defined in the AI Strategic Plan are:⁷²

Level 0: AI Not Used - No AI or autonomy integration in systems or processes.

Level 1: Insight - Human decision-making **assisted** by a machine. Al integration in systems is used for optimisation, operational guidance, or business process automation that would not affect plant safety/security and control.

Level 2: Collaboration - Human decision-making **augmented** by a machine. Al integration in systems, where algorithms make recommendations that could affect plant safety/security and control are vetted and carried out by a human decision-maker.

Level 3: Operation - Machine decision-making **supervised** by a human. Al and autonomy integration in systems, where algorithms make decisions and conduct operations with human oversight that could affect plant safety/security and control.

Level 4: Fully Autonomous - Machine decision-making with **no human intervention**. Fully autonomous Al in systems, where the algorithm is responsible for operation, control, and intelligent adaptation without reliance on human intervention or oversight that could affect plant safety/security and control.

As shown above, higher levels of AI autonomy reduce the need for human intervention or oversight. Consequently, these systems may require more stringent regulatory scrutiny, particularly if a system failure or malfunction could have serious implications for nuclear safety or nuclear security.

A parallel to this autonomy scale can be found in the automotive industry, where self-driving features in cars are becoming increasingly common. While current levels of automation do not yet achieve full autonomy, they offer various layers of driver support functions. The SAE J3016 standard was developed to provide a comprehensive framework for driving automation specifications and technical requirements. The standard details six levels of vehicle driving automation systems, outlining the extent to which these systems perform parts or the full dynamic driving task and the distribution of autonomous and shared functions between the driver and AI system at each level.⁷³ While automation is not the same as autonomy, parallels can be seen between the SAE and USNRC taxonomies, as demonstrated in Table 6.

SAE J3016 Levels of Automation	USNRC Levels of Autonomy
Level 0: No Driving Automation	Level 0: AI Not Used
Level 1: Driver Assistance Level 2: Partial Driving Automation Level 3: Conditional Driving Automation Level 4: High Driving Automation	Level 1: Insight Level 2: Collaboration Level 3: Operation
Level 5: Full Driving Automation	Level 4: Fully Autonomous

Table 6: Comparison of SAE J3016 Levels of Automation and USNRC Levels of Autonomy

72 Ibid., p. 1-4, Table 1: "Notional AI and Autonomy Levels in Commercial Nuclear Activities".

73 SAE International, "Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles", Standards J3016_202104, 30 April 2021. Available at: <u>https://www.sae.org/standards/content/j3016_202104/</u>.



While it is beyond the scope of this document to explain SAE J3016 in detail, the standard provides a practical model for describing the shared responsibilities between the intelligent machine and the human operator. As the nuclear industry embraces AI systems, such specification will likely be required to support both risk analysis and regulatory assessment.

AI Systems and Risk

When considering any technology integration into a nuclear facility, the level of risk needs to be assessed, specifically related to potential adverse impacts on nuclear safety and nuclear security systems. 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.⁷⁴ Regulations and regulatory guides are promulgated to support the licensees in meeting this expectation. To a large degree, such regulations and associated guides are technology-neutral, serving to establish performance goals, instead of specifying precise technology requirements. As an example, the USNRC Cyber Security Rule specifies:

- The licensee shall protect digital computer and communication systems and networks associated with:
 - Safety-related and important-to-safety functions;
 - Security functions;
 - Emergency preparedness functions, including off-site communications; and
 - Support systems and equipment which, if compromised, would adversely impact safety, security, or emergency preparedness functions.⁷⁵

While digital computer and communication systems are referred to here, the underlying technologies are not specified. Article (2) further indicates that such systems should be protected from cyber attacks that would adversely impact the systems' underlying functions. The technology-neutral language of the regulation allows for the broad application of digital technologies by the licensee with the stipulation that the licensee must show sufficient protection of said technologies from cyber attacks. The devil is in the details, however. Identifying the vulnerabilities of new digital technologies and associated mitigations can be difficult, representing challenges for both the licensee in security control implementation and for the regulator in compliance verification.

The United Kingdom's Office for Nuclear Regulation (UKONR) directly states its philosophy on technology neutrality, announcing that the current "goal-setting, outcome focused, risk-based regulatory framework is *technology neutral*".⁷⁶ Thus, the issue is not necessarily the use of AI technologies, but their application according to standard principles and a level of validation that nuclear safety and nuclear security functions are not or cannot be adversely impacted by their use. The key to regulatory acceptance when integrating AI applications into nuclear facilities relates to the specific function that AI is to perform and the associated risk to nuclear safety and nuclear security that would occur in the case of an AI system failure or maloperation. The UKONR establishes key regulatory expectations to this end:

- **Safety, security, and robustness:** Al systems should function in a robust, secure, and safe way throughout the Al lifecycle, and risks should be continually identified, assessed, and managed.
- Transparency and explainability: Al systems should be appropriately transparent and explainable.⁷⁷

74 Canadian Nuclear Safety Commission, UKONR, and USNRC, "Considerations for Developing Artificial Intelligence Systems in Nuclear Applications", September 2024. Available at: <u>https://onr.org.uk/media/03zl1osf/canukus_trilateral_ai_principles_paper_2024_08_28-final.pdf</u>.

75 USNRC, "§ 73.54 Protection of digital computer and communication systems and networks", 09 March 2021. Available at: <u>https://www.nrc.gov/reading-rm/doc-collections/cfr/part073/part073-0054.html</u>.

76 UKONR, "ONR policy: ONR's pro-innovation approach to Al regulation", Issue No. 1, April 2024, p. 5. Available at: https://www.onr.org.uk/media/v45dkpu2/onr-pro-innovation-approach-to-ai-regulation-paper.pdf.

77 Ibid., p. 7.



Subsequent to the USNRC AI Strategic Plan, nuclear regulators from the Canada Nuclear Safety Commission (CNSC), the UKONR, and the USNRC developed a joint report detailing considerations for implementing AI in nuclear facilities. The report discussed the risk dimensions of AI deployment in a nuclear facility. Using the USNRC 4 Levels of Autonomy, the report examined the significance of AI failure and adverse consequences to nuclear safety and nuclear security functions. Consideration was given to the level of autonomy and the ability of a human intervention to account for AI failure. One result of the report was the development of Figure 23.



Amount of AI Autonomy

Fig. 23: AI Applications Level of Autonomy and Categorisation of Risk of Failure⁷⁸

Referring to Figure 23, AI applications falling in areas 1 and 2, whose failure has minimal impact on nuclear safety and security, may be more readily applicable and accepted for use. While those in areas 3 and especially 4 will require stronger justification and rigor in verification to justify use.

The challenge of "Al verification" here is threefold. First, Al standards are in their infancy as the technology has faced rapid growth. Second, Al models often display non-deterministic, non-transparent, and non-intuitive behaviours. Thirdly, a significant body of operator experience does not exist to inform the use of Al applications in a nuclear facility. These three challenges give rise to regulatory caution and pause. Just as the operators of nuclear facilities seek to better understand the Al environment, so do the competent authorities.

78 CNSC, UKONR, and USNRC, "Considerations for Developing Artificial Intelligence Systems in Nuclear Applications", September 2024, p.3. Available at: <u>https://onr.org.uk/media/03z11osf/canukus_trilateral_ai_principles_paper_2024_08_28-final.pdf</u>





While AI integration into nuclear facilities is certain to progress, this will require significant effort and ongoing, close attention to ensuring nuclear safety and security.

Conclusions

Al technologies are currently in use in many applications in nuclear facilities, whether we realise it or not. The integration of such technologies, especially in business and social applications, can only be expected to increase, including in the nuclear sector, as the body of knowledge and operator experience with Al increases. Many of the concerns or reluctance related to broadening the use of Al in nuclear facilities centre around the lack of transparency in the Al applications, derived outputs, and the lack of workforce Al knowledge and expertise. However, similar concerns existed during the early integration of digital technologies at nuclear facilities. Systems that were previously mechanical, pneumatic, hydraulic, or otherwise analogue in nature were suddenly replaced by digital control elements, which increased complexity and lacked the observability of previous systems. Standards, regulation, and the state of practice for digital system integration has since matured, and in time, the same will almost certainly be said for Al integration.

Integration in the nuclear industry, however, will take tremendous effort and resources. Al adoption will likely be driven by operational factors and needs faced by the industry, for example, the need to:

- Reduce high O&M costs
- Reduce human errors
- · Improve workforce capacity building to use AI technologies
- Analyse overwhelming data streams and big data analysis



Some nuclear facilities, as demonstrated in the Nuclear Industry Survey responses, may choose not to adopt AI technologies due to a lack of sufficient need or to avoid any associated risks.

Nuclear regulation will be a consideration for AI implementation. Nuclear competent authorities recognise the need to develop their own capabilities and methodologies to adequately assess AI technologies to support risk-informed regulatory decision making.

One area to be closely assessed and defined is the expected and acceptable levels of performance of an integrated AI application in a nuclear facility. AI technologies have limitations, are not infallible, and will fail. Operators and competent authorities will need to assess the impact of failure that is considered acceptable. For example, in considering the case for autonomous vehicles, contemplate these figures. In 2021 police-reported traffic crashes in the United States were 6,102,936, in which 42,939 people were killed and an estimated 2,497,657 people suffered injuries.⁷⁹ If autonomous vehicles could reduce the number of deaths by almost 50 percent, resulting in 20,000 fatalities, would that be acceptable? If not, what number of fatalities would be acceptable? The point here is that the expectation for AI operation must be tempered, failure anticipated, and appropriately mitigated or accepted.

Advancements in AI technologies are ongoing with rapid growth in applications across all levels of industry and society. This will not stop. The key for the nuclear industry is the adoption of standards of practice and risk-informed methodologies to guide AI adoption in a safe, secure, and sustainable manner.

79 U.S. Department of Transportation, "Summary of Motor Vehicle Traffic Crashes", Traffic Safety Facts: 2021 Data, National Highway Traffic Safety Administration, DOT HS 813 515, October 2023.







Vienna Center for Disarmament and Non-Proliferation

The VCDNP is an international non-governmental organisation that conducts research, facilitates dialogue, and builds capacity on nuclear non-proliferation and disarmament.



vcdnp.org



info@vcdnp.org



@VCDNP