ADELARD

THE IMPACT OF AI/ML ON NUCLEAR REGULATION

24 Waterside
44–48 Wharf Road
London
N1 7UX

T +44 20 7832 5850
F +44 20 7832 5870
E office@adelard.com
W www.adelard.com

D/1321/165002/2 v3.0

Authors

[redacted]

Produced for

ONR contract number 117670V

THE IMPACT OF AI/ML ON NUCLEAR REGULATION

Executive Summary

The aim of this project is to advise ONR on the suitability of existing UK nuclear regulation to the application of artificial intelligence and machine learning (AI/ML), and to provide a route map towards gaining the benefits of these technologies in UK nuclear facilities.

AI/ML systems have huge potential to improve the safety of current and future nuclear plants and facilities through technologies such as autonomous vehicles (e.g., for the clean up of decommissioned sites), augmented intelligence (e.g., to alert an operator that safety systems may be under stress), and automation of simple operational tasks.

We begin by surveying the AI application landscape, and challenges in its assurance. The nature of ML algorithms frequently makes it impossible to walkthrough decision-making processes easily, or for systems to 'explain' their decisions. A significant challenge is therefore their opaque nature, which makes their actions difficult to interpret, their biases unclear and their malfunctions mysterious. Methods to adequately validate and understand these systems are still under development. Further challenges exist in security, novel development lifecycles and the availability of high quality standards.

We then analyse ONR's safety assessment principles to determine where AI/ML systems might be disruptive to existing regulatory approaches in the nuclear industry. We find that the principles are themselves strong, however the associated guidance will require interpretation and the development of additional rationale in order to be meaningfully applied to AI/ML systems.

Finally, we outline our recommendations to ONR and present a route map to AI/ML assurance.

We believe ONR can play a leadership role in AI/ML development by engaging in four processes:

- developing an AI regulatory framework, building upon the SAPs and associated guidance
- developing architectural approaches and a data strategy
- engaging with standards for AI/ML
- taking an active role to build capability within industry

Our route map presents a strategy to enable ONR to effectively, proportionately and fairly regulate this disruptive and powerful technology.

THE IMPACT OF AI/ML ON NUCLEAR REGULATION

Document control

Reference: D/1321/165002/2
Status: FINAL
Verified: [redacted]
Approved: [redacted]

Version	Review no./Issued	Date
v1.0	R/5041/165002/3	24 February, 2021
v2.0	R/5065/165002/4	31 March, 2021
v3.0	R/5101/165002/5	2 June, 2021

Distribution

[redacted] ONR, Bootle
[redacted] ONR, Bootle

THE IMPACT OF AI/ML ON NUCLEAR REGULATION

Contents

Figures

Tables

1 Introduction

The overall aim of this project is to advise ONR on the suitability of existing UK nuclear regulation with regards to the application and use of Machine Learning (ML) and Artificial Intelligence (AI) in operations affecting nuclear material. This report draws on specific research undertaken for ONR and other research being undertaken by Adelard on assuring autonomous systems.

AI has had a disruptive impact on a wide range of industries, and active research continues to develop AI systems for safety-critical sectors such as medicine, air travel and the nuclear industry, promising to improve safety and reduce human error. It is clear that AI has the potential to reduce risk in a nuclear environment, and it is essential that regulators prepare to understand and adapt to these incoming technologies.

ML is traditionally considered a subset of AI; ML is the study of computer algorithms that allow computer programs to automatically improve through experience and to generalise to novel settings, whilst AI is the science and engineering of making computers behave in ways that, until recently, we thought required human intelligence. A common machine learning technique is deep learning, which is particularly powerful for analysing unstructured and unlabelled data; as such, it is common in AI applications. These relationships are summarised in Figure 1.

A key application of AI is in autonomous systems. These are capable of independently performing actions to accomplish goals based on their knowledge of previous operations or by reference to external circumstances that are monitored and measured within the system. Although autonomy is a subset of AI, or is enabled by AI, not all uses of AI are autonomous. For example, many AI technologies intend to augment human decision-making, rather than replace it [1]. The scope of this paper considers ML, autonomous systems, and any other subsets of AI that may be applicable to operations affecting nuclear material.

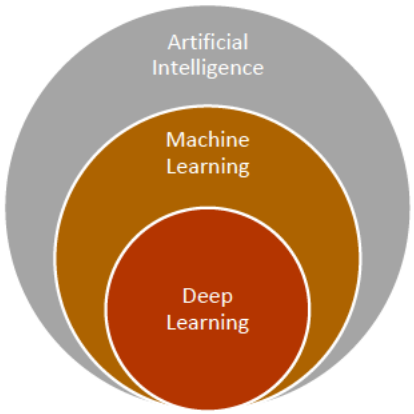

Figure 1: A Venn diagram of AI/ML techniques

On a historical note, in 1988 Adelard was contracted on behalf of the Commission of the European Communities to write a short paper warning the nuclear industry of the complexities of licensing AI systems in response to the initial boom of interest in the 1980s [34] for expert systems to support control room operations.

This report is structured as follows. In Section 2, we provide background and an overview of the AI/ML landscape, focusing on the types of systems available, nuclear-specific applications and difficulties in regulating and assuring these. In Section 3, we review ONR's safety assessment principles to identify areas

that may require clarification, additional guidance, or may be otherwise effected by AI/ML assurance. Our recommendations for a route map towards gaining the benefits of AI/ML systems are outlined in Section 4. Finally, our discussion is summarised in Section 5.

2 Opportunities and challenges in AI/ML

2.1 What is machine learning and artificial intelligence?

Artificial Intelligence (AI) is the effect of a technology performing activities that have traditionally only been associated with humans. The activities that are considered 'intelligent' may change over time as machines become increasingly capable. They typically perceive their environment and perform complex reasoning to establish the optimal way to meet their goals. Modern AI technologies include speech recognition, self-driving cars and chess engines. These intelligence effects can be driven by a range of computational methods; possibly the most powerful approach is that of machine learning, in which a computer algorithm is able to improve its performance through experience.

Machine Learning (ML) is a broad term covering many different methods of generating a computer algorithm to provide predictions or decisions, using selected training data and adaptation. Typical examples include image classifiers, which filter image data and predict location and type of objects in an image using a neural network, and data mining, which looks for trends and patterns in large data sets. ML technologies can range from probabilistic Bayesian inference to artificial deep learning neural networks.

ML generally requires large amounts of training data to converge on a solution. This training data should be representative of the deployment task of the ML in terms of the types and population densities of those types. Training data may be tightly curated for ground truth, such as labelled images with bounding boxes, or more loosely managed, such as partially populated decision trees or data generated via randomised simulations.

Broadly speaking there are four types of ML:

- unsupervised learning – using training data with no labels or ordering
- supervised learning – using training data which has been curated in detail
- semi-supervised learning – using training data which has been loosely curated (e.g. into groups but with limited info on how to interpret such as no bounding boxes)
- reinforcement learning – given an end goal can the ML come up with a solution that generalises (e.g., using simulation to generate its own scenarios)

As well as training data, validation data may be used during training to check current performance and is often sampled from the training data. Completely independent test data is used to assess performance. The test data should represent the items of concern and may have different population profiles than the training/validation data. Other forms of testing include simulations, ranging from fully virtualised settings and environments to in situ real-life testing, and combinations thereof.

In terms of training, ML systems are usually deployed in one of two ways:

- as a continuously learning system in which data encountered during operation serves to continuously update the ML algorithm
- as a pre-trained system, in which the learning takes place once only, before deployment, and the algorithm does not develop during operation

The former technology allows systems to adapt to new or novel surroundings more easily, and respond to operator feedback to improve performance. This comes with additional assurance burdens, however, as the behaviour of the system can change over time and may become unsafe. In this report, we have assumed the use of pre-trained systems, unless otherwise stated.

2.2 The AI/ML technology and application landscape

Key AI areas of potential relevance to the nuclear industry include autonomous systems (e.g., for the clean-up of decommissioned sites), augmented intelligence (e.g., to alert an operator that safety systems may be under stress), and control of simple operational tasks that might otherwise have been assigned to a human operator. In this section, we survey the current state of the art in AI/ML systems and what technology is likely to become commercially available in the coming years.

2.2.1 The Hype Cycle

The Gartner Hype Cycle (shown in Figure 2) shows the current level of maturity and adoption of different AI products, and their readiness for real-world application. Most innovations progress through a pattern of inflated expectations and disillusionment before becoming sufficiently mature to reach real-world productivity. Gartner argues this behaviour is seen across the spectrum of new technologies, and therefore, the curve can be useful to predict the future impact of a technology. Note that technologies may move at different rates across the cycle, but as the technology becomes more widely adopted, it will more commonly start to appear in more conventional products.

The Hype Cycle helps us understand what technologies are 'on the doorstep', which are a long way away, and which may never live up to current expectations. It is clear that whilst applications in augmented and decision intelligence systems are potentially on the horizon, general artificial intelligence capabilities are a long way off and may never live up to the original expectations.

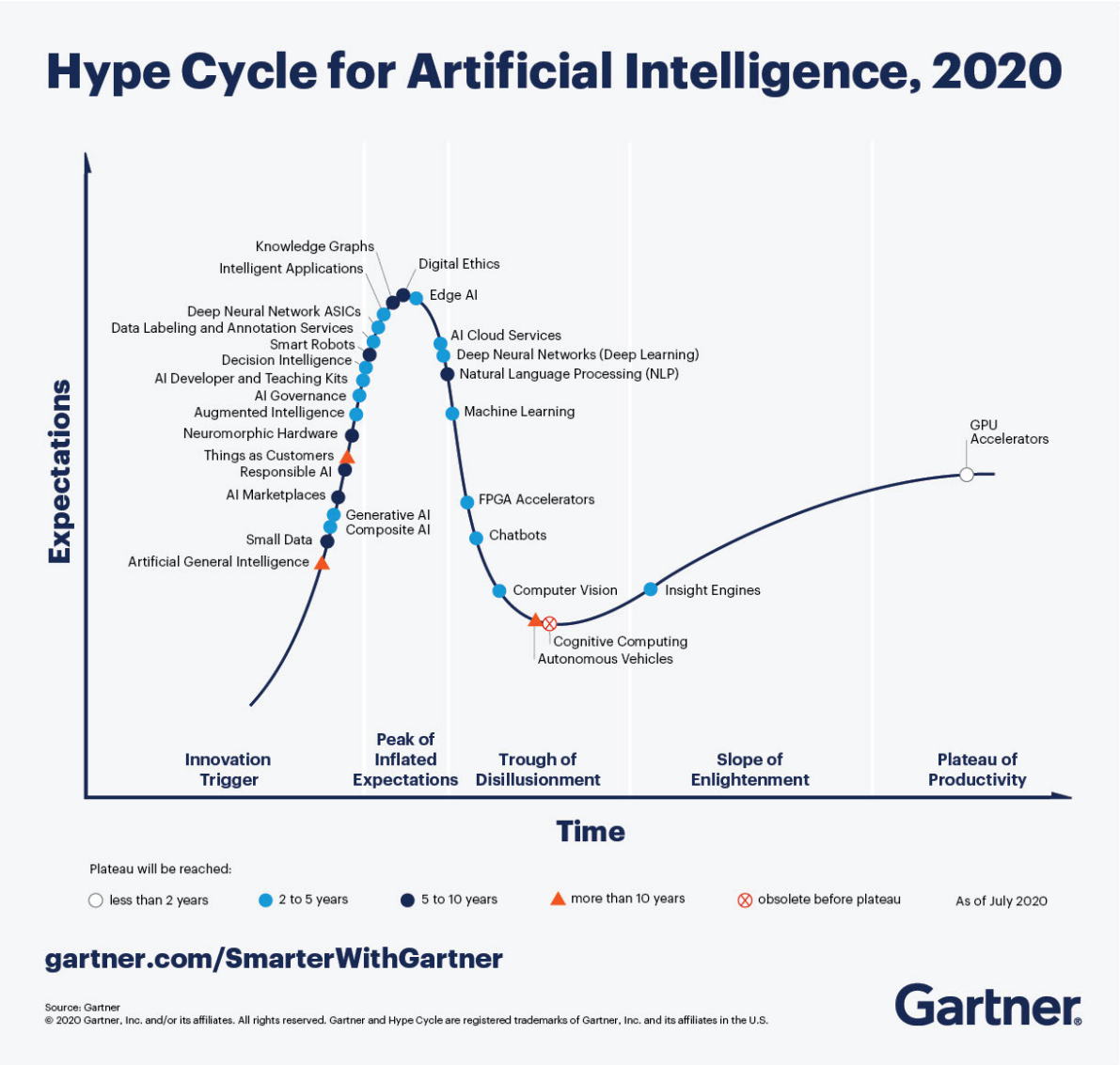

Figure 2: The Gartner Hype Cycle for AI

Gartner recommends three factors when considering when to adopt an innovative technology:

1. how valuable the innovation may be to the organisation
2. where the innovation is on the Hype Cycle
3. how tolerant of investment risk the organisation is

Conservative organisations minimise risk by adopting new technologies only once they have reached the Plateau of Productivity, where they are mature enough to have well-defined criteria for assessing provider viability. The nuclear industry comprises a wide range of applications with a range of levels of conservatism required. Regulators should therefore be aware that some technologies under development may be further from fruition than current expectations may suggest.

Some AI technologies that may appear in the short term, with nuclear sector importance, could include

- augmented intelligence and decision intelligence systems which may have implications for operators and control of systems
- custom ASIC and FPGA accelerators are allowing AI to be placed into embedded devices

In the longer term, smart robots and autonomous vehicles could play a role in future decommissioning processes. The challenges that need overcoming to reach these goals are outlined in Section 2.3.

2.2.2 Processor evolution and new forms of computing

Oftentimes, AI/ML-based systems require high performance computing for both training and operation. This has an impact on the architectures utilised to support such resources, especially for those applications requiring lower costs (for example, embedded devices).

The architectural structure of AI/ML systems is continuously changing and developing. Developments in chip design, such as tensor processing units (TPU), ASICs and FPGA accelerators combining with new optimisation techniques, such as quantisation aware training, have significantly reduced the computational cost of machine learning algorithms. This has opened the door to bringing AI on to embedded devices and allowing AI accelerators on to chips. Eventually these technologies will start working their way into the supply chains of companies; particularly as development toolkits and platforms are reducing the entry-level knowledge required to generate AI algorithms.

There is now a general trend to develop specific AI/ML processors with the main chip suppliers creating special purpose chipsets dedicated with AI/ML in mind. In particular, Intel intend to create specific chip sets for autonomous systems where monitor architectures (which aim to constrain the ML algorithm in a system to a limited, safe set of states) are pushed into the silicon as much as possible. Furthermore, NVidia have been bringing Tensor Core accelerators into their GPUs with their recent A100 GPU released in 2020 specifically designed for AI, data analytics and high performance computing. These cards are currently aimed at data centres, but their A6000 GPU is already being marketed for modern servers and workstations for AI workloads.

There is also the possibility of more radical changes to computing chip design. Brain-like computing architectures, conceived in the late 1980s, are progressing under the banner of neuromorphic computing. Neuromorphic computing is aiming to remove the rigid, narrowly defined, rules-based problems that current generation AI/ML systems are limited by. This could create AI systems that are far more flexible and less brittle to novel situations where the AI system may have little prior context. Current neuromorphic systems that attempt to mimic the way the brain processes information today rely on FPGAs, CPUs and GPUs. Conversely, neuromorphic processors intend to simulate neurons and synapses to mimic the activities that occur in the human brain.

IBM has developed a neuromorphic chip, named TrueNorth [2]. It has 4,096 cores, each with 256 neurons with 256 synapses each. The microprocessor has a power density orders of magnitude lower than that of a conventional von Neumann processor. It has been built into a cluster with 16 million neurons and 4 billion synapses. Another example is the Loihi research chip from Intel – a fifth-generation self-learning neuromorphic chip – which was introduced in November 2017 [3]. These chips support probabilistic computing, which aim to create algorithmic approaches to dealing with the uncertainty, ambiguity, and contradiction in the natural world.

It is clear that the technology landscape in the AI/ML sector is rapidly developing. In order to use this technology safely, constant monitoring of the state of the industry and accepted best practices will be needed.

2.2.3 Nuclear-specific applications

The RAIN (Robotics and Artificial Intelligence for Nuclear) project has been created for the purpose of developing robotic and AI technology to solve challenges faced by the nuclear industry. The initiative aims to "offer major opportunities for improving productivity and significantly reducing risks to human health" given the issues of radiation levels and extremely harsh conditions that restrain humans from appropriately accessing and operating some nuclear facilities [4]. Developed and ongoing robotics and AI systems include

- remote inspection robots for difficult terrain and underwater settings, including the first ever autonomous radiometric survey of a former alpha laboratory on the Sellafield Ltd site
- development of pipe inspection robots
- robot generated VR for scene modelling using deep learning
- 3D mapping with handheld sensors at walking pace

An IAEA report [5] looking at machine learning in the nuclear domain provides additional examples of how the different ML types might be applied to different repetitive tasks for inspectors. Unsupervised learning could be used to identify anomalies that lie outside of "normal" sensor data boundaries, such as monitoring liquid solutions and their density. An example of supervised learning may be to count fuel rod assemblies as part of international safeguarding.

Various advances in AI technologies will have either a direct or indirect effect on the nuclear industry as well, such as:

- Autonomous vehicles and intelligent robots to help the clean-up of decommissioned sites or entering areas at high risk for humans to perform tasks.
- Augmented intelligence and decision intelligence systems to alert operators that safety systems may be showing the first signs of being under stress by recognising subtle and complex patterns from plant sensors. Even going further and reducing the load on the operator by taking control over some of the simple operational tasks that still require a human level of intelligence.
- New hardware advances in deep neural net ASICs, such as tensor processing units (TPUs) and FPGA accelerators are allowing AI to be placed in embedded devices and microcontrollers. This may lead to even 'smarter' digital devices with enhanced diagnostics and more complex functionality.
- AI requires vast amounts of data but advances in cloud computing and decentralised distributive processing is removing these barriers, which could allow for much richer data collection and analysis providing better predictions for maintenance or knowing when systems may be on the edge before failing.

2.3 Current challenges in assuring AI/ML systems

The development of AI technologies presents a challenge to regulators. It is often difficult to fully specify the requirements for AI/ML products, as they often involve responding to complex and diverse inputs. A significant challenge is the "black box" nature of AI/ML products, which make their actions difficult to interpret, their biases unclear and their malfunctions mysterious. Methods to fully validate and classify these devices are still in development.

For systems handling complex and diverse inputs, it is infeasible to achieve adequate coverage using simple black box testing. This is for two main reasons:

- An unfeasibly large number of tests is required to adequately sample input space.
- The classification algorithm in ML systems is often a poorly behaved function of the inputs (i.e. there are often cliff-edges and artefacts), so it is never possible to be confident that defective behaviour has been accounted for. This is illustrated in Figure 3.

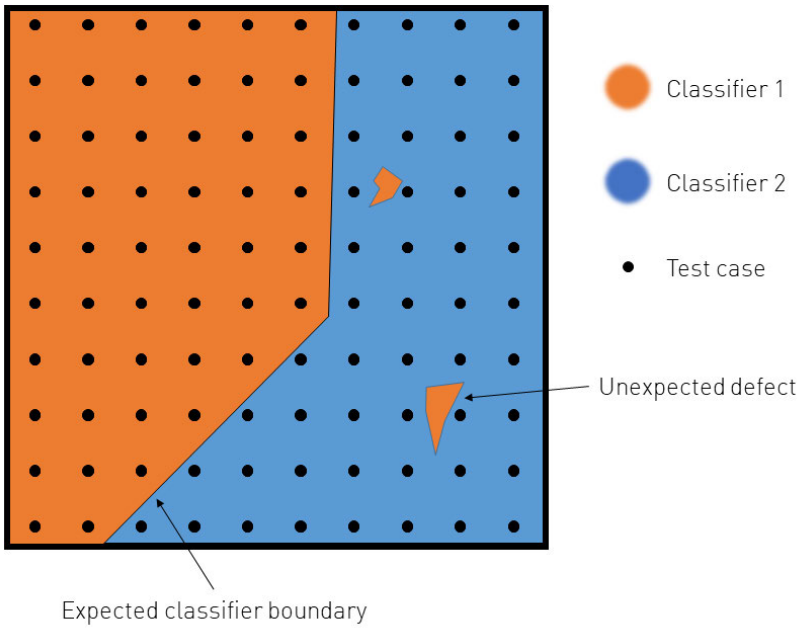

Figure 3: Even robust testing may fail to detect defects, due to the nature of ML classification algorithms. In this example, two defects sit between test cases, so are undetected.

2.3.1 Assurance approaches and safety cases

Many certification approaches for safety-critical applications rely on showing detailed conformance to standards. However, where there are novel technologies or applications, there is an increasing trend to propose outcome, property, and behaviour focused assurance approaches. The variety of "frameworks" from industry [6] [7] [8] [9] all envisage an outcome-based approach and structured safety or assurance cases. The key question is how to implement such an approach.

Our work in TIGARS [10] focused on the assurance of first-generation autonomous vehicles (AVs), or more generally robotic & autonomous systems (RASs), currently being deployed and how existing approaches for assurance need to change to address current and future autonomous systems. The project provided a cross-sector and international (UK-Japan) perspective.

We produced a number of TIGARS Topic Notes (TTNs) to support the development and evaluation of autonomous vehicles. These TTNS are publicly available [10] and will be published by the Centre for the

Protection of National Infrastructure (CPNI) soon. The TTNs address the challenges faced in the current landscape. Additionally, we discuss potential solutions and recommendations proposed by a varied set of literature as well as preliminary research that we have carried out. The accompanying TTNs are

- Assurance – Overview and Issues
- Resilience and Safety Requirements
- Open Systems Perspective
- Formal Verification and Static Analysis of ML Systems
- Simulation and Dynamic Testing
- Defence in Depth and Diversity
- Security-Informed Safety Analysis
- Standards and Guidelines

We noted that the autonomous systems field is international and has a wide variety of players of differing maturity. Some entrants are unfamiliar with classical safety engineering, yet have expertise related to AI and ML-based systems. Others are mature and familiar with classical assurance approaches but lack a grasp on the challenges autonomy brings. This is likely to be the case in the nuclear application of AI/ML as well. Our recommendations for building safety cases for these AI/ML systems are outlined in more detail in Section 4.2.3.

2.3.2 AI/ML tool chain and development lifecycles

The AI/ML algorithm development lifecycle tends to be a more iterative approach compared to a more traditional software development lifecycle. Often the AI model and data sets require refining as the model matures over many training and evaluation cycles until the model is able to meet its functional and performance requirements. The testing and evaluation of models may have multiple stages that test and evaluate different aspects of the algorithm, such as black box testing, simulation, and case studies and trials.

Figure 4 shows a simplified typical ML component development cycle. Higher-level device requirements, such as function, performance and reliability, will flow down to requirements on the AI/ML component. A deployed ML algorithm will be the outcome of the development process; however, this algorithm will probably require updating and maintenance throughout its operational lifetime. Any updates affecting the ML algorithm itself, including updates to the ML model, the training data or anything else requiring additional training of the model, will require repeating the majority of the testing and verification process, given the black box nature of much of this testing.

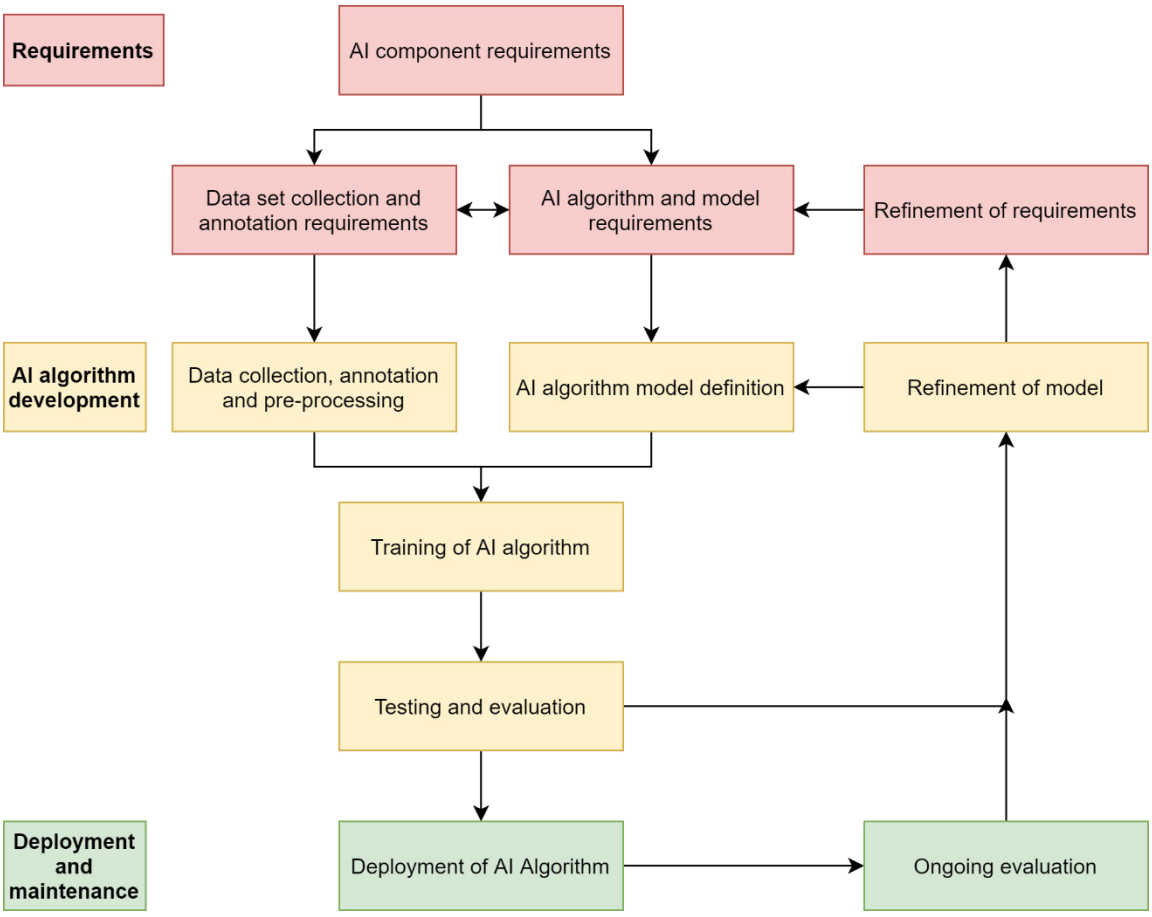

Figure 4: ML component development lifecycle

2.3.2.1 Platforms and tools

Autonomous systems contain more than just the AI/ML components. Traditional systems exist throughout, either separate from the AI/ML functionality or in support of it as a platform to develop and run the AI/ML component.

The supporting software is paramount to building AI/ML models, whether it be through using pre-existing third party toolkits and libraries or in-house platforms to manage and implement AI algorithms and their large data sets; supporting tools are involved with designing the architecture of the AI model/algorithm, training and refining the model, and validating it through simulation and testing benches. Platform tools are also needed for the run time environment, preparing and feeding input data to the AI algorithm and parsing the results.

Typical vulnerabilities faced by the non-AI/ML supporting systems must still be considered against all the attributes under consideration, such as reliability, performance, cyber-attacks and availability.

To improve confidence in the supporting platform, static and dynamic analysis can be performed on supporting software to look for potential issues and vulnerabilities that could impact the performance or

functionality of the code. These analyses do semantic and syntactic checks on the source and object code, and prevent runtime errors when the program is executing.

Furthermore, researchers have shown that overflow errors within supporting software can propagate and affect the functionality of an ML model, as they identified a vulnerability in a robotics system where a Not a Number (NaN) code error could cause uncontrolled acceleration [11].

Runtime issues such as overflow/underflow and access to data out of bounds can directly impact on the performance and outputs of an ML element such as a neural network, as it may perform a series of matrix multiplications on edges and nodes. Also, it may be more difficult to identify which parts of the software are affected by the error, and hence, the potential impact.

Software bugs in the supporting platform that was involved in the training and evaluating of the AI Algorithm could have an impact on the performance of the model, and fixing them may impact the performance of the trained model. If the model should be re-trained due to the initial flaw being fixed is something that should also be investigated.

2.3.2.2 Data sets

Data sets utilised to train AI/ML systems have a proven history of being riddled with implicit biases, due to either incompleteness or lack of inclusivity in training data sets. These biases within the training data often lead to indirect prejudice and discrimination in the algorithm. When selecting or creating training datasets, it must be the case that diversity, potential bias, ethics, privacy, and fairness are all considered. This is discussed in more detail in Section 3.8.

2.3.3 Standards and guidance

The assurance of many software-based components deployed in safety-critical systems often takes on a standards-based approach, as standards are a key tool for establishing ALARP arguments, and more specifically, have a role in defining production excellence for computer-based safety systems.

However, in the case of AI/ML-based components, it is not always possible to solely rely on standards, as they are lagging behind the state-of-the-art advancements and deployment of ML systems. Nevertheless, standards are important in defining and promulgating good practice and shared terminology and concepts.

In the past two years, standardisation bodies have begun concentrating their efforts on the assurance of AI/ML systems, with a few produced outputs available. However, the landscape is changing quickly, and a number of activities are expected to be completed within the next couple of years.

In the progress interim report, we outlined the landscape of international standards and guidance currently being developed to assure the implementation and deployment of AI, ML, and autonomous systems that may be relevant to operations affecting nuclear material (see Section A.1). Standardisation of AI/ML systems is an active area and resulting outputs have increased significantly in the last two years with many publications to be scheduled between now and the next several years.

In Appendix A we provide an in-depth landscape review of the standards presently available:

- ISO/IEC TR 24028:2020 Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence
- IEEE 7010-2020 Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being
- ANSI/UL 4600 Standard for Safety for the Evaluation of Autonomous Products [6]

- High-Level Expert Group on AI in European Commission — Ethics Guidelines for Trustworthy Artificial Intelligence [12]
- SASWG's Safety Assurance Objectives for Autonomous Systems [13]
- Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims [14]

AI/ML technologies and methodologies are developing quickly and it is not clear how much the standardisation will address the key issues specific to AI/ML. An egregious example is ISO 22737, a standard for requirements and testing of Low-Speed Automated Driving (LSAD) systems, which excludes ML-based sensors and only requires approximately five tests. This loose level of rigour is not compatible with the stringent requirements expected of the assurance of safety-critical systems, especially within the nuclear domain.

Other standards such UL4600 provide a comprehensive list of objectives and seek to address the whole of the AV system lifecycle. However, this repeats much of the same grounds as existing safety standards. One advantage of UL4600 might be its more outcome focused approach and its use of pitfalls, that could be used as a source of defeaters. In later work, we intend to further investigate its utility to augment PE or to define best practices, but our initial observation is that it addresses development aspects by citing IEC 61508 and IEC 26262. However, the objectives and clauses outlined may be useful in reviewing safety cases.

In general, upcoming standards would need to address the complete AI/ML lifecycle (in particular, for safety-critical systems), and demonstrate guidance on how to address the range of new vulnerabilities AI/ML systems present in terms of both the data utilised, and the model itself.

2.3.4 Security

The security of nuclear plants and facilities is a very broad topic encompassing possible safety incidents that could be caused maliciously, whether through primary technical means, social engineering, direct attack or, as is more likely, some combination of all of these. The responsibility for safeguarding nuclear materials brings additional requirements for the confidentiality and integrity of sensitive nuclear information. Since AI/ML systems bring with them unique security issues, it is particularly important to develop a strong security framework for their assurance.

Security and safety have developed as distinct disciplines, with their own regulation, standards, culture and engineering. Security requirements for nuclear power plants are addressed in the Security Assessment Principles (SyAPs) [15] and specifically for computer-based systems important to safety (CBSIS) in TAG 46 Appendix 6. IEC 62859 specifies the requirements for coordinating safety and cyber security, but not in a way fully integrated with safety. It mentions possible failure modes and consequences of cybersecurity features in safety functions, but they are still considered separate activities in the lifecycle, with security being an add-on that has to be justified because of the increased complexity it adds to the design. It does not consider the impact of functional safety requirements on security and the possible hazardous consequences from an attack or intrusion of the system. The impact of integrating security when developing a safety demonstration of a smart device is discussed in [16].

Given the links throughout the lifecycle between security and safety, it is inappropriate to treat the two activities completely separately, as there is a growing understanding of their close interconnection. For example, it is no longer acceptable to assume that a safety system is immune from malware because it is built using bespoke hardware and software, or that it cannot be attacked because it is separated from the outside world by an "air gap". In reality, the existence of the air gap is often a myth (see [17] [18] [19]).

Furthermore, autonomous systems rely on data and software with uncertain provenance and are not designed for high integrity applications. A safety justification, or safety case, is incomplete and unconvincing without a consideration of the impact of security threats.

The impact of cyber security issues is exacerbated by the increasing sophistication of attackers, the commoditisation of low-end attacks, and the increasing vulnerabilities of digital systems as well as their connectivity – both designed and inadvertent [19].

The following areas are particularly significant from a security perspective and need more scrutiny in a security-informed justification of a safety system [20] [21].

1.	Integration and interaction of requirements, e.g. of safety, with security and resilience supported by security-informed hazard analysis techniques.
2.	Supply chain integrity, e.g. mitigating the risks of devices being supplied compromised or having egregious vulnerabilities.
3.	Malicious events post-deployment that will also change in nature and scope as the threat environment changes, and a corresponding need to consider prevention (e.g. implementing a risk-based patching policy) but also recovery and resilience.
4.	Weakening of security controls as the capabilities of the attacker and technology change. This may have a major impact on the proposed lifetime of installed equipment and design for refurbishment and change.
5.	Reduced lifetime of installed equipment as there is a weakening of security controls as attackers' capabilities and technologies change.
6.	Threats to the effectiveness and independence of safety barriers and defence in depth.
7.	Design changes to address user interactions, training, configuration, and software vulnerabilities and patching. These might lead to additional functional requirements for security controls.
8.	Possible exploitation of the device/service to attack itself or other systems and the need for confidentiality of design and deployment information.
9.	The trustworthiness and provenance of the evidence offered.

Table 1: Security-informed safety issues

Some of these issues could be particularly problematic for the nuclear industry:

- the need for increased scrutiny of supply changes for digital systems and components given the low importance of nuclear industry to the AI/ML industry
- the difficulty of assessing the effectiveness of defence in depth and architectural measures
- the trade-offs between the risks and disruption of patching with the risks of the vulnerabilities being exploited

To address these issues there is a need to integrate security into safety analyses so that the interactions and trade-offs that are necessary can be considered. For example, at the requirements stage, we might need to consider the security aspects of the information flow policy when a plant is under attack, or if degraded plant conditions impact the safety. Another type of issue that we might need to consider at the architecture level is whether a highly critical third party component has sufficient security provenance given

its supply chain. Safety assessment involves building trust with the supply chain, visiting their factories and assessing their culture: these are all aspects highly relevant to security as well as safety.

Overall, the need for security-informed safety impacts the complete safety lifecycle, from policy and requirements, through implementation and operation, to decommissioning and disposal.

2.3.4.1 Security-informed risk analysis

One of the key topics in security-informed safety is the impact of security on risk assessments covering the whole lifecycle. One of the TIGARS Topic Notes [20] yet to be published by the Centre for the Protection of National Infrastructure (CPNI) focuses on this area.

The TTN guidance states that security concerns could have an impact on

1. the system boundaries
2. what systems could potentially affect safety
3. the stakeholders involved
4. the validity of design safety assumptions

Therefore, care must be taken during the analysis to account for security concerns as well as safety. Table 2 summarises a seven-step risk assessment process.

Step	Brief description
Step 1 – Establish system context and scope of assessment	Describe the system to be assessed and its relationship with other systems and the environment. Identify the services provided by the system and the system assets. Agree the scope of and motivation for the assessment and identify the stakeholders and their communication needs. Identify the type of decisions being supported by the assessment.
Step 2 – Configure risk assessment	Identify any existing analyses, e.g. safety cases, business continuity assessments that provide details of the system, the impact of failure and the mitigations that are in place. Characterise the maturity of the systems or project and the key uncertainties. Ensure that the risk assessment is focused on the kinds of threats that are of concern. Define possible threat sources and identify potential threat scenarios. Refine generic capability and impact levels for the systems being assessed. Identify risk criteria. Refine and focus system models in the light of the threat scenarios and existing analyses to ensure that they are at the right level of detail for an effective security-informed risk analysis.
Step 3 – Analyse policy interactions	Undertake an analysis of policy issues considering interactions between safety requirements and security policies. Resolve any conflicts, show that the trade-offs are satisfactory and document the decisions made.

Step	Brief description
Step 4 – Preliminary risk analysis	Undertake architecture-based risk analysis, identifying potential hazards and consequences and relevant vulnerabilities and causes together with any intrinsic mitigations and controls. Consider doubts and uncertainties, data and evidence needs. Identify intrinsic and engineered defence in depth and resilience.
Step 5 – Identify specific attack scenarios	Refine preliminary risk analysis to identify specific attack scenarios. Focus on large consequence events and differences with respect to the existing system.
Step 6 – Focused risk analysis	Prioritise attack scenarios according to the capabilities required and the potential consequences of the attack. As with the previous step, the focus is on large consequence events and differences with respect to the existing system.
Step 7 – Finalise risk assessment	Finalise risk assessment by reviewing implications and options arising from focused risk analysis. Review defence in depth and undertake sensitivity and uncertainty analysis. Consider whether the design threat assumptions are appropriate. Identify additional mitigations and controls.

Table 2: 7-step security-informed safety risk assessment

Both security and safety perspectives are needed to assess the likelihood of vulnerabilities being exploited and the effectiveness of the overall architecture and other security controls in limiting the impact of exploits. There is a variety of initiatives to integrate security into hazard analyses. We have been using security- (or cyber-) informed Hazard analysis and operability studies ("Hazops") [21] to assess architectures of industrial systems [22], and adapt this well-known approach for performing a safety hazard analysis in a systematic fashion [23], analysing the deviations of data flows and values between different interconnections in the system. To account for security in a security-informed Hazops, additional security guidewords are added and an enhanced multidisciplinary team (system safety and security experts) is used.

2.3.4.2 Security of AI/ML-based systems

AI/ML components add another level of complexity onto security risks. An AI algorithm is often created using a supporting platform and requires large amounts of data to train; these increase the available attack surface for security attacks on the products if the overall development lifecycle is not secure.

The security risk of using publicly available third party AI/ML platforms should be assessed, as it opens up the possibility of attacks appearing in the supply chain; however, it may be too expensive and unpractical to build an in-house bespoke AI/ML framework from scratch. Therefore, mitigations in these situations can be taken to limit the security risk:

- using diversity and defence in depth in the overall system to limit the impact of supply chains issues
- developing components of the AI/ML framework in-house where feasible – the model/AI algorithm architecture, training process and collected data sets
- only using data sets from trusted sources – if possible run a profiling analysis on the data sets before use to ensure that it meets the expectations
- keep third party libraries and components up to date
- re-train and optimise any public base model with private data

Special care must be taken in AI/ML systems to guarantee privacy and data protection throughout a system's lifecycle. ML transferability properties, in combination with the ability to query ML models, can reveal information regarding the ML model and parameters, compromising Intellectual Property (IP). Additionally, it has been shown that sensitive personal data, if used in training, can be extracted from the ML model outputs. Restrictions and limitations on querying of the AI/ML system should be considered to protect user privacy and data.

Adversaries could aim to influence and exploit the collection and processing of data, corrupt the model, and manipulate the resulting outputs [24]. Notoriously, researchers in [25] have forced models to make wrong predictions by computing what are now known as adversarial examples. These are examples that produce perturbations that are very slight and often indistinguishable to humans, yet are sufficient to change the model's prediction to one that is incorrect, such as that shown in Figure 5.

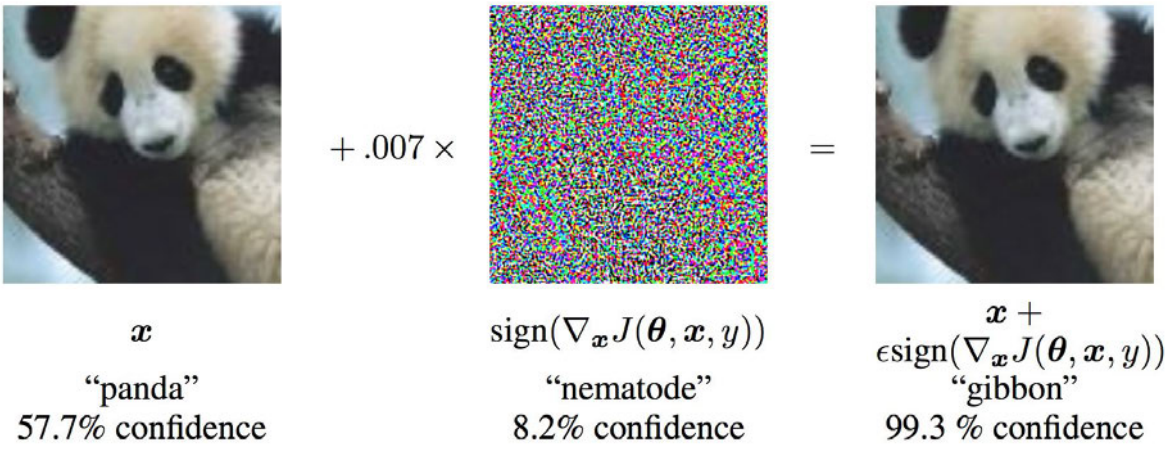

Figure 5: The fast gradient sign method introduced by [25], but also a potentially misleading example [26]

Intriguingly, the actual attacks presented as in Figure 5 do not work in practice as they do not take into account the integer representation of pixels [26]. The adversarial perturbations are fragile, and quantisation destroys its ability to delude an image classifier. However, [26] presents a new quantisation mechanism which preserves the adversarial nature of the perturbation.

Adversaries might be capable of manipulating the model inputs to affect its output, thus reducing the robustness, accuracy, availability, and integrity of the overall behaviour of the system. It should be noted that these are complex attacks that require direct access to the ML model and supporting platform to be able to query and read outputs from the model and also inject the perturbations to the input stream.

Nevertheless, issues like this demonstrate the fragility of these types of systems. Small changes to the input data can have large effects, and ML models may be unable to deal with unexpected inputs, for example adversarial t-shirts (see Figure 6) or unexpected animals crossing a road.

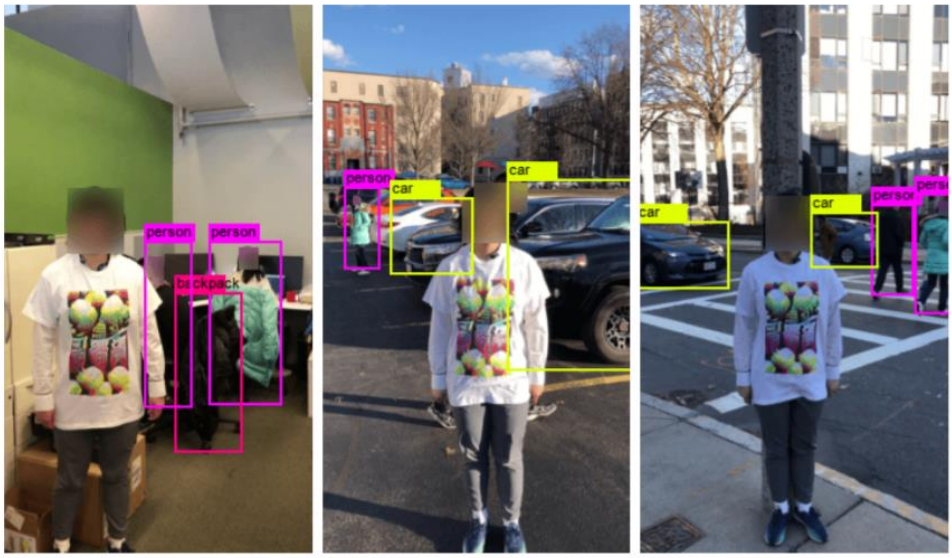

Figure 6: An adversarial t-shirt can avoid detection by the YOLOv2 system [27]

This fragility is a generic property of these types of systems and not an isolated result. In [28][29] the authors discuss that this is in the nature of AI/ML systems that they are fragile and not just an implementation issue. It is not just associated with vision systems, but is a general feature of ML. In [28] they discuss the use of common ML classification algorithm techniques, such as cross-entropy loss functions and how the low-rank features of the training data are responsible for the effectiveness of adversarial inputs, like the classic panda recognition example in Figure 5, and the fragility of these types of algorithm. They result in creating very poor decision boundary margins meaning a small perturbation in the input space can have large effect on the classification result.

The authors do present potential solutions for these types of image classifiers with differential training that allows the feature mapping to remain trainable, like these current classifiers, but increasing the input space of classification decisions; this results in these algorithms being fooled less by adversarial inputs and requiring larger perturbation in the input space to change the feature mapping sufficiently enough to affect the output classification (improving robustness). These examples highlight the need for safeguards, defence in depth and diversity as these systems on their own may not be able to achieve the reliability of more traditional digital devices. It also highlights the sophisticated research focus on attacking AI/ML systems.

3 Suitability of existing guidance for AI/ML regulation

In this section, we review the nuclear regulatory approach for its applicability to AI/ML assurance with the aim to highlight where the regulations may need to be clarified or updated for AI/ML systems.

Nuclear regulations are encapsulated in the UK nuclear safety assessment principles (SAPs), which place additional guidance within a collection of technical assessment guides (TAGs), along with a large body of supporting standards and guidelines. These act as guidance for ONR inspectors.

The highest level of guidance in the SAPs are the fundamental principles, which are considered the foundation for the subsequent safety and radioactive waste management principles. Here, we first use these fundamental principles to highlight key themes where AI/ML products may be disruptive to the assurance process. We then investigate these areas in detail by studying the engineering, leadership and management and safety case principles as outlined in the SAPs, as well as the accompanying TAGs.

Our detailed recommendations for how these clarifications can be made are given in Section 4.

3.1 Fundamental principles

The fundamental principles (FP) cover a wide scope regarding high standards of nuclear safety. From these we have identified three that are particularly relevant to the use of AI/ML. These precipitate several themes, outlined in Table 3, which will be discussed in more detail in following subsections.

FP	Topic	Principle	Issues to be explored with AI/ML-based systems	Key themes
1	Responsibility for safety	The prime responsibility for safety must rest with the person or organisation responsible for the facilities and activities that give rise to radiation risks.	Whether the use of AI/ML-based systems would undermine the responsibility of duty holders. Where systems are autonomous, the delegation of authority must be addressed.	Human factors
3	Optimisation of protection	Protection must be optimised to provide the highest level of safety that is reasonably practicable.	Optimisation may be achieved with the use AI/ML systems. For example, a claim might be made that adding an AI/ML adviser would improve safety performance of control operators. However, how would their reliability be demonstrated?	Determining reliability Security Production excellence
4	Safety assessment	Duty holders must demonstrate effective understanding and control of the hazards posed by a site or facility through a comprehensive and systematic process of safety assessment.	For AI/ML systems that have a role in mitigating or possible initiating hazards, what does it mean for a duty holder to "understand" the AI/ML operations and how might a systematic safety assessment be carried out for this technology given its complexity, adaptability and lack of predictability?	Understanding and explainable AI Safety cases Data management

Table 3: Relevant SAPs, fundamental principles and potential concerns with regard to the use of AI/ML

From Table 3 it is clear that the application of AI/ML will have a large and wide-ranging impact on safety assurance in the nuclear industry. The key themes of complex human factors, reliability, security,

explainability and building safety cases are discussed individually in detail in the following subsections, in which we search the more detailed engineering and leadership and management principles, and their associated TAGs, for relevant guidance.

3.2 Understanding and explainable AI

As highlighted in FP4, ONR requires licensees to *understand* what their safety systems can and cannot do, and the consequences of their failure. ML and AI systems, however, are often opaque to human understanding, acting as a 'black box'; that is, they cannot be understood, even in principle. This is a result of the convoluted learning process, making it impossible to 'walk through' decisions in the same way as is possible with traditional software. As such, these requirements require clarification for AI/ML assurance; to what extent (and in what sense) are licensees expected to be able to understand their systems' behaviour?

The difficulty in understanding AI/ML systems can be attributed to two main causes:

- The nature of ML algorithms frequently makes it impossible to walkthrough the decision making process easily.
- AI/ML systems are typically deployed to solve complex problems with huge input spaces (for example, combining data from multiple complex sensors), for which it is not practical to write a complete and unambiguous specification. This makes it impossible to verify the system's behaviour for every possible scenario it may encounter.

It is therefore useful to make a distinction between understanding *why* a product took an action, and understanding the product enough to be able to *predict* what it will do on a given input. The first is the goal of explainable AI, a relatively recent direction in AI development. The second could theoretically be understood by testing every possible input. Neither are currently feasible for most commercial AI products.

The SAPs reference understanding throughout the management and engineering principles. Some key clauses are discussed in Table 4.

Reference	Relevant guidance	Comment
MS.2	Experts must have "a detailed and up-to-date understanding of the safety of its facilities and their design, operation and safety cases".	In the context of AI/ML components, this may imply fully explainable AI is required, however it may be possible to partially meet this requirement by an analysis of the type of failure, and an investigation to indicate whether it was caused by insufficient training data, a software failure or another cause. Clarification is required.
MS.4	"A learning organisation should … understand the reasons for differences between actual and intended outcomes."	
EHA.5	The hazard analysis should allow for "understand[ing] the behaviour of the facility in response to the hazard".	

Reference	Relevant guidance	Comment
EHF.7	"User interfaces should…provide sufficient, unambiguous information for the operator to maintain situational awareness"	It is essential a human operator can understand and interpret the decisions made by the AI/ML product in order to best react to its decisions and make use of its outputs. This includes understanding the reliability of the AI/ML product in the current situation, and the reasoning for any decisions made by the AI/ML product. This is of particular importance to AI/ML systems that may be making decisions under operator supervision, where the huge volumes of data processing performed can easily swamp user interfaces, unless designed carefully. This should be strengthened and clarified.
The Regulatory Assessment of Safety Cases	"Safety analysis requires an extensive understanding of the facility, both in the present and foreseeable future, its behaviour in a variety of conditions and experience of failures… It also requires an understanding of how people and organisations may affect safety."	It is difficult to understand with confidence how an AI/ML component would behave under arbitrary conditions. This could be partially met by testing the model under a wide range of common or significant conditions.

Table 4: Key SAPs clauses about understanding

In order to have systems that are adequately safe, we need to understand the reasons why they are safe and we need to take responsibility for that safety. We need to understand what might go wrong, what has gone wrong in the past, whether it is relevant and how it might be mitigated. As such, clarified guidance is required on this topic.

3.2.1 Intelligent customer capabilities

TAG49 (Licensee Core Safety and Intelligent Customer Capabilities) outlines the responsibilities of the licensee with respect to knowledge and understanding of the plant design and safety case. Licensees must maintain a 'core safety capability' which includes capabilities as an 'Intelligent Customer': "The licensee must be in control of activities on its site, understand the hazards associated with its activities and how to control them, and have sufficient competent resource within the licensee organisation to be an 'Intelligent Customer' for any work it commissions externally."

The SAPs define being an intelligent customer as: "The capability of an organisation to understand where and when work is needed; specify what needs to be done; understand and set suitable standards; supervise and control the work; and review, evaluate and accept the work carried out on its behalf."

Both of these definitions seem compatible with having 'black box' systems as part of the plant, and so no further clarification is required.

3.3 Determining reliability

Being able to clearly and verifiably assess the reliability of a computer-based system is of clear importance in safety-critical systems. This process for AI/ML systems may be qualitatively different to that for traditional software, due to the high-dimensionality of the input space they operate in (i.e., the system combines many and complex input data), and the complex lifecycles required to develop them.

The ONR SAPs reflect this across the document; some key guidance is outlined in Table 5. Additionally, TAG46 outlines the two-legged approach to reliability assurance, requiring both evidence of production excellence (PE) and independent testing (ICBMs). In this section we first discuss how PE and ICBMs may be applied to AI/ML systems. We then discuss how demonstrating reliability may be different for AI/ML compared to traditional software.

Reference	Relevant guidance	Comment
ERL.1 & ERL.2	"Adequate reliability and availability should be demonstrated by suitable analysis and data... Evidence should be provided to demonstrate the adequacy of these measures."	What is the best way to establish reliability for an AI/ML system?
ESR.2	"The reliability, accuracy, stability, response time, range and, where appropriate, the readability of instrumentation, should be adequate for it to deliver its safety functions."	
ESS.27	"Where the system reliability is significantly dependent upon the performance of computer software, compliance with appropriate standards and practices throughout the software development lifecycle should be established in order to provide assurance of the final design."	What standards and practices are appropriate for AI/ML development?
ESS.27	Production excellence must include an "independent assessment of the comprehensive testing programme"	How can a comprehensive testing programme be performed for AI/ML?

Table 5: Reliability clauses in the SAPs

Demonstrating the reliability of AI/ML systems is often a complex task – this is due to the high-dimensional input space in which they operate (combining many highly detailed inputs), meaning it is very difficult to adequately sample the possible inputs and determine if all relevant edge cases have been tested and measured. Edge cases are inputs that only occur at an extreme operating parameter, for example operating in extreme darkness. A further problem is that AI/ML systems often behave probabilistically rather than deterministically – for example an image classifier might be expected to correctly identify a traffic light in an image 80% of the time. Of course, all computer-based systems are strictly deterministic, however AI/ML systems work in such a large and complex input space it is usually more helpful to consider how the system reacts to groups of inputs (for example, images of traffic lights), in which case its behaviour appears probabilistic.

It is clear that, whilst the SAP's requirements for substantiated reliability claims are appropriate for AI/ML regulation, the types of evidence required to back up these claims will be substantially different from traditional computer-based systems. For example, extrapolating behaviour and performance from testing

to real-world scenarios is less simple. Testing is needed not only to evaluate products on a bounding scenario, but also to challenge the system with a sufficient set of representative scenarios to build confidence in their effectiveness. These claims will likely require more extensive testing than traditional deterministic claims, as statistical approaches are needed. Appropriate metrics to measure the accuracy of ML algorithms are essential.

In Appendix B, we give an overview of some common probabilistic metrics used to evaluate an AI/ML system's performance. These are also summarised in Table 6. It is important to note that a reliable system is not necessarily safe – if the system works in a complex dynamic environment, it could carry out its function perfectly as specified but still cause an accident.

Classification type	Example	Helpful metrics
Binary classifier	Alarm trigger	ROC AUC, Precision, Recall, F_1, F_β
Object detection	Exploratory robot for mapping locations of items	Intersection over Union, AP^k, $AP^{0.5}$, AR^d
Object tracking	Tracking payload location for crane monitoring	Mostly tracked, mostly lost, identity switches, fragmentations, multiple object tracking accuracy, multiple object tracking precision

Table 6: Summary of metrics discussed in Appendix B

3.4 The two-legged approach

The SAPs place requirements on the way that software-based systems are justified for use in safety-critical parts of a nuclear installation. The so-called "two-legged" approach breaks the assessment into two parts [30]:

- Justification of production excellence (PE) – "a demonstration of excellence in all aspects of production from the initial specification through to the finally commissioned system". It should include the following elements: technical design practice consistent with current accepted standards; implementation of a modern standard quality management system; and application of a comprehensive testing programme formulated to check every system function.
- Independent confidence building measures (ICBMs) – "Independent 'confidence-building' should provide an independent and thorough assessment of the safety system's fitness for purpose". This should include the following elements: complete, and preferably diverse, checking of the finally validated production software by a team, including independent product checking that provides a searching analysis of the final system; independent checking of the design and production processes, including the activities undertaken to confirm the realisation of the design intent; and independent assessment of the comprehensive testing programme covering the full scope of the test activities.

3.4.1 The meaning of production excellence

One reading of the SAPs is to interpret "all aspects of production" to mean that the *process* of production has to be excellent. Another interpretation of "excellent" is that the products of that process have to be

excellent too (arguing that how can a process be excellent if it does not produce excellent products?). (Also, if being pedantic it does not say that testing should be successful).

TAG 46 clarifies this to some extent. In Section 5.2.4 "Production excellence is achieved through the application of relevant good practice at the time of system development to avoid errors, detect and remove those not avoided and provide in-built system tolerance to those not detected". Yet it also focuses on weaknesses in the process, not on the product: "Should the production excellence assessment identify weaknesses in the production process, compensating activities should be applied to address them. The type of compensating activities will depend on, and should be targeted at, the specific weaknesses found." Note that the TAG allows defects to be present in the software.

TAG 46 Section 5.4 introduces the idea that PE may be about software defects: "Production excellence: a demonstration of excellence in the production of the computer based system important to safety to minimise the likelihood of the introduction of latent systematic faults in the software development process; and..."

So does PE provide a "thorough assessment of the safety system's fitness for purpose", as the ICBMs do, or does it provide an assessment of the production process and the fitness for purpose of the artefacts produced (is this a good design, is the test plan a good example of a test plan)?

3.4.2 The meanings of ICBMs

The two-legged metaphor used to describe ICBMs suggests two reasons for believing in the suitability of something and that either would be sufficient but a more robust case can be made with both. Indeed the SAPs require that "Independent 'confidence-building' should provide a thorough assessment of the safety system's fitness for purpose" suggesting just that, although in practice this does not seem to be the case, especially for safety class 2 and 3 applications.

The TAG describes independent confidence building measures as "confidence gained through the application of independently conducted, diverse from production, techniques and methods used to assess the system software and hardware." This might suggest they are partial and rather less than a thorough assessment of the safety system's fitness for purpose.

On the other hand, in 5.29 it states "The confidence building leg provides an independent and thorough 'reasonably practicable' assessment of the safety systems' fitness for purpose".

3.4.3 Combining PE and ICBMs

One interpretation for the different roles of PE and ICBMs is that

- PE establishes software is fit for purpose (there might have been some areas of weakness, but these will have been compensated)
- ICBMs also establish software is fit for purpose

We then have more confidence than before because PE and ICBMs are dissimilar, in part because the ICBMs have been designed without taking into account the PE results. The nature and extent of the diversity will impact the gain in confidence (whether to use forced diversity or not also is a design issue for the assurance. It might be better for the diversity to be chosen in knowledge of the PE findings rather than designed in ignorance of the PE).

A second interpretation is

- PE establishes software is fit for purpose.
- ICBMs tackles where the system is judged independently (but not in ignorance of the PE) to be vulnerable, where it thinks the PE case is weak and seeks other areas which might be problematic. It seeks to maximise the gain in confidence but is only partial.

A third interpretation is that

- PE establishes the software production process and artefacts produced are excellent
- ICBMs establishes the software is fit for purpose

We can summarise these interpretations as

- Two assessments that by reasons of accidental diversity increase our confidence in fitness for purpose, both providing adequate confidence on their own.
- One assessment that demonstrates fitness for purpose with another targeted assessment designed to maximise confidence in the system.
- One assessment that provides confidence in the artefacts and process followed. Another assessment demonstrates fitness for purpose.

Indeed there are further variants, for example exploring what "fitness for purpose" means.

It is also not clear quite what the claims are being made for the independence aspects of the ICBMs – are they addressing lack of diversity in tools (so possible common errors), mindset issues and missing things, organisational failure, doubts in the underlying theories? Or do they address all of these?

3.4.4 Techniques and a graded approach

The lack of specific claims associated with the PE and ICBMs make it hard to assess the importance of any issues or gaps that are found and also to adjust for different levels of criticality. The SAPs point to graded approaches found in standards where typically the use of techniques is varied across the levels of the safety function supported by the device (e.g. as in IEC 61508). The techniques can be considered in two classes: those that provide evidence of system properties (e.g. timing, reliability) and those that address defects or vulnerabilities in the code and provide indirect evidence of system properties.

For smart devices, the Cinif SING project mapped techniques to properties as does Annex V of the IAEA Guide [31]. One reason this is difficult is that there is a lack of theories or detailed argumentation linking the results of the techniques to the properties that are of interest at a computer system level. This makes developing and justifying a graded approach particularly hard.

This lack of detail also needs to be taken into account when considering the validity and application of standards and a graded approach. The experts in the standards group should have had sufficient experience to be able to judge the impact of a technique. Nevertheless, in doing so they were dealing with making a correlation between the systems they have assessed and the impact of using techniques. What would be a sufficient evidence base for their judgement? The lack of explicit validation of standards and maintaining the rationale for them makes development of an equivalent approach for new technologies like AI/ML difficult.

There is therefore a need to bridge the gap between a set of techniques and the impact of using them has on the system property of interest: to do this we need to formulate claims about what the technique demonstrates and then seek theories that connect the claims to the system level claim of interest. In CAE

terms we have a direct claim, e.g. from static analysis, and then need some theory to do the lifting up to claim a certain pfd (via a substitution block).

One should also consider that standards and regulations can make recommendations that are, on average, good policy, which lead to better results over a population of systems. There is then a need for specific systems to show what has been achieved. This applies to the standard as a whole, as well as to specific techniques. For example, diversity in software development can be shown on average to increase reliability, but whether a specific system has improved, and by how much, needs specific evaluation.

3.4.5 Application to AI/ML-based systems

In considering the applicability of the SAPs and TAG46 to AI/ML we have been hampered by the lack of rationale for these. Specifically, this makes it difficult to formulate the claims we would make for having done PE, the claims we make for undertaking a particular technique and those for independence. In addition, in applying and interpreting the PE and ICBMs requirements and grading them with different technologies and criticalities one needs to know their impact and connection with the system properties.

The SAPs are not linked to higher level principles such as FP3, which might have provided an overarching principle to appeal to when the specific PE and ICBMs ones seem hard to interpret.

While some of these issues could be clarified by a more precise analysis e.g., using the CAE concepts, they are not all just issues of clarification. There are gaps in the theories and evidence base as well.

Nevertheless, despite these concerns about the detailed interpretation of the SAPs, one can identify some specific issues that the application of them would raise:

- How to define PE for AI/ML systems when the claims that can be derived from the use of these standards is unclear as they are generally relatively new. The status and role of standards is discussed in more detail in Section 2.3.3.
- How, in the application of ICBMs, the requirement for a "complete, and preferably diverse, checking of the finally validated production software" could be interpreted and achieved given the nature of the AI/ML-based systems and the high-dimensionality input space of many AI/ML systems.
- What tools and techniques could provide the "searching analysis of the final system"? What analysis tools (static analysis, model checking of properties, simulations) could be deployed to provide this?
- How would "independent checking of the design and production processes, including the activities undertaken to confirm the realisation of the design intent" be applied to the AI/ML lifecycle?
- How might "application of a comprehensive testing programme formulated to check every system function" be interpreted and achieved?
- What the applicability and interpretation of statistical testing for AI/ML systems may entail (note that an element of this may be part of the AI/ML development lifecycle which typically involves a range of simulation approaches).
- How is checking of specifications undertaken if it is embedded in the training data?

In addition, there are three generic topics that are important in consideration of ICBMs and PE: deterministic vs probabilistic approaches, black box issues and understanding.

The UK approach is often described as deterministic with probabilistic evaluations to confirm the deterministic design (see SOCS [32]), with considerable variations internationally on how probabilistic risk assessments inform and are involved in the plant design stages. However, even the deterministic approach has uncertainties associated with the analysis and judgements, and so the corresponding claim will have some uncertainty associated with it. The probabilistic approach provides a framework for evaluating and

exploring this uncertainty, providing it can cope with epistemic issues. In the use of AI/ML systems they will be uncertain in their performance, just as people are, and metrics of uncertainty and benchmarks will be needed to evaluate them (see the discussion in Section 3.2.1 and A.2.6).

There is also the perennial issue of how much access to detail is required for the PE/ICBMs. In the past, the emphasis has been on evaluation of source code and reliance on proven-in-use arguments for the processors. With the use of special purpose devices such as FPGAs, this has changed somewhat and will be further complicated as more safety algorithms become pushed into the silicon. There is need for validated principles that can be applied across technologies covering when access to information is required. This so-called Black Box problem would benefit from specific principles on this topic as well as from pulling together elements of understanding (see Section 3.2) and PE/ICBM issues.

A first stage in addressing these would be to develop more clarity in the two-legged approach and address in more detail its potential application to AI/ML:

- Model the different interpretations of the two-legged approach with a range of parameters (e.g., efficacy of confidence building, predominance of difference types of error, different sources of doubt) and see under which circumstances they give better regulatory outcomes. Consider the application to training data and AI/ML algorithms.
- Consider the impact of moving to a property focused, rather than technique focused approach. Identify gaps in knowledge of the techniques' efficacies and how and whether these could be addressed. Consider the application to training data and AI/ML algorithms.
- Consider how uncertainty and confidence are addressed and how regulation might deal with non-deterministic systems and the uncertainties associated with AI/ML.
- Develop a principled approach advising when access to detail is required – the Black Box issue – and how lack of access might be mitigated. This should be integrated with elaborating the principles for understanding and integration into wider SAPs with a rationale.

These last two issues have a wider impact than just the two-legged approach.

In parallel, research could be initiated to develop the requirements and capabilities need for architecture and defence in depth analysis, as well as techniques necessary to demonstrate safety properties. This is elaborated in the overall route map and discussed in Section 3.9

3.5 Human factors

3.5.1 Integrating ML with human operators and staff

This section considers issues where current tasks requiring human factors (HF) analysis are being either replaced or enhanced by AI/ML systems but there is still always a human in the loop. In a Section 3.5.2 we consider this on a different scale – where the human is complete replaced.

As more tasks are managed by AI/ML systems, there must be consideration of how these can be successfully integrated with human operators/maintainers and other staff who use them as part of their tasks. Superficially it could be assumed this is no different from current human factors analyses, but there are new (less predictable) ways in which AI/ML can fail compared to traditional computer systems or other humans, and they can potentially change behaviour slowly over time (with reinforcement learning or due to

a changing external environment). Maintenance may also change behaviour in unexpected ways[1]. The way in which data is presented to the operator could also change – depending on how the ML "explains" its results. There may also be a different response from the human, for example resistance to introduction to more automation, or low level of trust in the technology leading to more interventions. Alternatively, if the human trusts the technology too much (so called "automation bias") they may fail to notice issues with its performance which could have a safety impact.

One potential benefit of introducing more automation – if the complexities of shared decision-making are addressed – is reduced workload and hence stress and fatigue could be reduced. However, again, there may be a negative reaction from the human operators, who feel less in control of an AI/ML system and hence more stressed and less able to cope with novel situations. This could lead to poor decision making and interventions. There might also be wider social issues of concern over jobs losses and automation.

We note that we can consider differing levels and types of involvement of a human to with a system with ML, for example

- continuous or regular interaction with a monitoring system
- restricted interactions, such as a robot which is programmed with high level goals to go from A to B and check C and the ML makes decisions on route as to how to achieve this
- as an end user who is impacted by the outputs but doesn't have any interaction with the inputs, such as an algorithm which automatically changes heating levels depending on weather and time of year/day
- as an uninvolved third party who is indirectly (and possibly adversely) affected e.g. a person caused harm as a robot knocks something over or fails to stop

All of these situations may impact existing practice, and assumptions about workload and vigilance levels. Of course, any workload changes would need to be assessed and the situations where it provides benefits and those which might increase risks identified and analysed to demonstrate whether more automation is actually providing material benefits. The "ironies of automation" [33], in which increased automation can lead to operator expertise decreasing and operators being less able to cope with difficult situations, informs nuclear industry practices.

There will also be the need to deal with adaptive systems and those with subtle adaptations that might not be visible to the user (a trivial example is reprogramming the text predictive system in your phone, based on a user's frequently used phrases). Being subtly recalibrated might be serious for a nuclear alarm or sensing system as it could undermine the safety analysis and human factor assumptions.

The three main TAGs associated with Human Factors are the following:

- Human Factors Integration (NS-TAST-GD-058)
- Human Reliability Analysis (NS-TAST-GD-063)
- Human Machine Interface (HMI) (NS-TAST-GD-059)

[1] We assume that the ML can be demonstrated to be of sufficient integrity and performance for the task required, otherwise it would not be installed. However, its failure/fault behaviour may still be different to conventional software systems.

The Human Factors Integration (HFI) TAG is principally intended to provide guidance to aid inspectors in the application of the following SAPs:

Number	Principle	Comment
EHF.1	A systematic approach to integrating human factors within the design, assessment and management of systems and processes should be applied throughout the facility's lifecycle.	There may be changes needed to ensure systematic integration with the new technology at whatever lifecycle stages it is used. The HFI plan would need to ensure integration with ML was adequately covered, and risks understood.
MS.2	The organisation should have the capability to secure and maintain the safety of its undertakings.	In order to adhere to this principle the organisation may need to expand roles and responsibilities and consider the technical competence required at many different levels. Ensuring that all staff or persons potentially impacted by the introduction of ML have the required skills.

Table 7: SAPs supporting HFI TAG

Complementary to HFI, the Human Reliability Analysis (HRA) TAG aims to identify and analyse all human actions and administrative controls that are necessary for safety. The guidance mainly focuses on the interpretation and application of principle EHF.5 and EHF.10:

Number	Principle	Comment
EHF.5	Proportionate analysis should be carried out of all tasks important to safety and used to justify the effective delivery of the safety functions to which they contribute.	There are a lot of issues relating to human reliability analyses which may change depending on how the ML is integrated into the overall lifecycle for the facility. Task analysis may be different. For example, if new ML is used to replace a task, what are emergent issues from this? How many tasks are disrupted by the introduction of the technology? Are there fewer staff? Are they differently trained? Demands in terms of perception, user interface design, decision making and actions may all be different. Another consideration is also the speed at which a human could react. ESS.8 states that no human intervention should be required for fast acting faults (30 minutes). It should be considered whether this is still a reasonable assumption for a system with ML.

Number	Principle	Comment
EHF10	Human reliability analysis should identify and analyse all human actions and administrative controls that are necessary for safety.	Human reliability analysis methods are likely to be challenged by the introduction with ML, if the ways in which the human could be unreliable changes i.e. there are new potential failures or changes in reliability. They could either have too much or too little trust in the new technology. Additionally, stress levels may increase if the operator did not feel in control of the new technology. This would in turn affect the likelihood of intervention and hence reliability. The difficulty of designing and assessing the co-operation between AI/ML systems and humans is exemplified in the difficulty in justifying handing back control in autonomous vehicles when the automation fails. As noted previously, activities such as calibration and maintenance of ML and its dependent systems may require expertise about how the functionality could alter and be prone to human error. Reliability data should ideally be derived from experience data which may not exist in large amounts for some of these tasks. This would challenge the validity of the analysis. If simulators are used for HRA their limitations should be identified, many ML simulators are not designed for the safety critical environment.

Table 8: SAPs supporting HRA TAG

Of particular importance is the consideration that humans play a key role in the safe and efficient operation of nuclear facilities, as they typically take on the role of an "operator". If an AI/ML system is to either replace or assist an "operator" for a task, Human Machine Interfaces (HMI) must also be re-considered and adapted to support the use of AI/ML systems to control the plant and manage nuclear safety. The HMI TAG specifically notes that "operators contribute to a plant's defence-in-depth hierarchy in a number of ways including the prevention and control of abnormal operation, detection of failure, control of faults within the design basis and accident/emergency response". Thus, human-based safety claims must be adapted to include AI/ML operations affecting nuclear material and other safety issues. The TAG considers the following principles:

Number	Principle	Comment
ESS.3	Monitoring of plant safety: Adequate provisions should be made to enable the monitoring of the facility state in relation to safety and to enable the taking of any necessary safety actions during normal operational, fault, accident and severe accident conditions.	The introduction of new monitoring systems, and potentially newly automated activities would need to be demonstrated to be at least as good as existing human performed tasks, and probably better (to justify the changes required). This may be a complex argument. Typical benefits of automation include improvements in response time, improved concentration for repetitive tedious tasks, more detail than a human is capable of. However, if it cannot be demonstrated that that these are consistent, reliable, and supportive of humans still involved then it may be difficult to make the case.
ESR.1	Provision in control rooms and other locations: Suitable and sufficient safety-related system control and instrumentation should be available to the facility operator in a central control room, and as necessary at appropriate secondary control or monitoring locations.	The SAPs guidance for this principle indicates C&I should be available in normal, fault conditions and severe accidents. It may be the case that certain ML would be suitable for only some of those situations. This may not simply be avoidance of ML for more severe conditions, but if used in severe accident conditions it could remove the risk from one or more humans. As noted, this would need justification in the safety case.
ESR.7	Communications systems: Adequate communications systems should be provided to enable information and instructions to be transmitted between locations on and, where necessary, off the site. The systems should provide robust means of communication during normal operations, fault conditions and severe accidents.	As more automation and ML is introduced, the nature of the communications systems may change. For example, more of the communication messages may be automated and more monitoring introduced of those communications.
ESR.8	Monitoring of radioactive material: Instrumentation should be provided to detect the leak or escape of radioactive material from its designated location and then to monitor its location and quantity.	The use of ML for this type of task, or to improve the measurements required, may in some cases provide a clear benefit over sending in a human. New HMI may be required to control such systems. However, that instrumentation must be reliable for use.

Number	Principle	Comment
EHF.7	User interfaces: Suitable and sufficient user interfaces should be provided at appropriate locations to provide effective monitoring and control of the facility in normal operations, faults and accident conditions.	The guidance supporting this principle provides some useful performance requirements on any product including ML, some of which may be very challenging to demonstrate as being *consistently* provided by ML. It should • provide sufficient, unambiguous information • provide a conspicuously early warning • support effective diagnosis • enable the operator to determine and execute appropriate actions

Table 9: SAPs supporting HMI TAG

For completeness, we have also reviewed the remaining human factors principles (in order of appearance) in the SAPs. Key comments are given in Table 10.

Number	Principle	Comment
EHF2	When designing systems, dependence on human action to maintain and recover a stable, safe state should be minimised. The allocation of safety actions between humans and engineered structures, systems or components should be substantiated.	This principle still holds assuming that the AI/ML can perform this task to the dependability/availability/reliability required to maintain safety. However, there may be a trade-off argument around which is most appropriate: • an automatic system with ML and no human action • an automatic system with ML and human oversight as safety backup (which is possibly against the principal but could be a way to gain confidence in ML unless there are so few incidents that this provides no useful information) • keeping the human action (again, this is against the principle) even if automatic system with ML available as the human is deemed more reliable/safe overall or simply because the ML can't be demonstrated to be equivalent or better If there an inherent assumption about human performance then it could be carried over to the ML as a benchmark of improvement.

Number	Principle	Comment
EHF3	A systematic approach should be taken to identify human actions that can impact safety for all permitted operating modes and all fault and accident conditions identified in the safety case, including severe accidents.	This principle still holds but not if there are new actions (e.g. maintenance and calibration) that a human could perform which have an impact on ML model distributions (such as changing training data or re-calibrating an input sensor to an NN). This should be identified as part of the safety analysis. Other tasks such as safeguards which may change based on ML support should also be considered.
EHF6	Workspaces in which operations (including maintenance activities) are conducted should be designed to support reliable task performance. The design should take account of the physical and psychological characteristics of the intended users and the impact of environmental factors.	As discussed previously, it's possible that the introduction of different types of machine learning based systems will bring different challenges. For example, they may control the environment (temperature controls), use physical space (robots), or have different interfaces depending on explainability.
EHF8	A systematic approach to the identification and delivery of personnel competence should be applied.	The competence required to interact with ML-based systems may be changed or have a different basic skill set. Any user or third party (assuming they fall under the responsibility of a licensee e.g. are on site for any reason) that can impact the ML in a way that relates to safety may need some retraining. The timescales and required time to ensure current competency (assuming the ML itself may change or there are sufficient changes in the environment that its performance has changed) could be significant. FA.13 specifies that possible human errors leading to faults must be covered in the probabilistic safety analysis.

Table 10: Remaining human factors related comments

In conclusion, our review of all three of the TAGs showed that the principles were in general unchallenged, but the underpinning analyses and processes currently used to ensure they are met have many uncertainties.

The principal that was potentially challenged was EHF2: "*When designing systems, dependence on human action to maintain and recover a stable, safe state should be minimised. The allocation of safety actions between humans and engineered structures, systems or components should be substantiated*". There may be a hidden underlying assumption that the engineered systems and controls are more reliable and trustworthy than

the human actions. There is also ongoing discussion in the AI community about the level of AI/autonomy that is really required or feasible. It may be that clarifying this principle, or its supporting guidance, should be considered to express this trade-off between minimising reliance on a human, but also appropriate reliance on AI/ML.

Other challenges (either to guidance or supporting tools and techniques) could be broadly grouped as covering

- competency – a wide range of competency training may be required for operators, maintainers and to all staff simply sharing space with new autonomous systems
- undermining existing models – models of human behaviour are challenged, particularly as new systems are first introduced, meaning reliability and stress analyses may be uncertain and hard to justify
- re-allocations of tasking – there may need to be fundamental changes in how tasks are grouped and allocated, further challenging models
- human machine interfaces – limitations in ML technology in terms of reporting and explainability may alter the ability to provide robust HMIs

Finding the right balance between humans and machines is a topic the nuclear industry has led on. Other industries are likely to be earlier adopters of AI/ML, (e.g., in semi-autonomous applications in road and defence) and any initiatives should leverage that work.

3.5.2 Anthropomorphic viewpoint – machines as people

At very high levels of autonomy and sophistication, approaching AI systems like digital ones may become less helpful. In the case that AI advances to the extent that it can learn a wide range of behaviours, take complex decisions independently and be taught and retrained in a wide range of contexts, it may be more helpful to view its regulation and behaviour through an anthropomorphic lens. This could be particularly significant when an AI system directly replaces a human with a specific qualification or management role. In this case we can shift from thinking of the dependability of these systems as complex automatic gadgets (for instance, how we can apply DO178C or IEC 61508 to them) to whether and how their assurance addresses the principles of understanding, explanation, challenge, and learning.

In this section we review the extreme case where ML has completely replaced human operators, but their function is still viewed as a series of "tasks". There is a philosophical question - at what point does the replaced human task stop being considered a replacement and just a normal part of the engineering design? However, for the sake of this report we have assumed the human factors principles may be considered. For example, task analysis might be a useful way to model how an automated system will behave. A workload model needs to consider number of tasks allocated and ability of the ML system to perform them consistently over a long period. A naïve interpretation of fatigue could include battery running out of power and assume long term functional fatigue is not likely (lack of concentration). However, overload of tasks could be an issue for the technology as the system might drift in situational awareness. Additionally, an awareness that the workload is about to exceed expected parameters would be needed (see ESS.13 which requires notification that a limiting condition has been exceeded).

The following are some speculative examples of full replacement of human tasks:

- Fully autonomous monitoring systems with ML – this would be a replacement of a repetitive task. We could assume the monitoring system is largely passive and its only action would be to indicate a problem, but that should be reliable. Workload problems translate to become performance issues but if the task is simple this may not be an issue.

- Fully autonomous inspection robot – the robot would be complex with multiple tasks to be modelled. Workload may be an issue as might prioritisation of tasks in a crisis situation, as well as the need to recognise that it is in one.
- An ML-based decision making or advisory system – an ML-based decision tree could perform this type of task, for example based on some data points (certain alarms combined with sensor readings) decisions could be made about actions to be taken. ML can be used for this type of task where its performance is an improvement over a conventional system (general due to speed and sparse or subtle input data).

Number	Principle	Comment
EHF1	A systematic approach to integrating human factors within the design, assessment and management of systems and processes should be applied throughout the facility's lifecycle.	This still seems appropriate as a high level principle, ensuring autonomous systems are integrated in a systematic way.
EHF2	When designing systems, dependence on human action to maintain and recover a stable, safe state should be minimised. The allocation of safety actions between humans and engineered structures, systems or components should be substantiated.	It is unclear exactly how to apply this principle in a very autonomous situation. See the discussion in Table 10.
EHF3	A systematic approach should be taken to identify human actions that can impact safety for all permitted operating modes and all fault and accident conditions identified in the safety case, including severe accidents.	This still seems appropriate as a high level principle, but would likely be covered by hazard analysis. One particular issue of note might be self-calibrating or reinforcement learning systems.
EHF4	Administrative controls needed to keep the facility within its operating rules for normal operation or return the facility back to normal operations should be systematically identified.	Our decision tree example may fall under this principle. In this situation the validity of the decision tree used for training, and the trained ML would both need strong validation.
EHF5	Proportionate analysis should be carried out of all tasks important to safety and used to justify the effective delivery of the safety functions to which they contribute.	This principle would still apply but may be supported by engineering (V&V) evidence instead. There is a requirement to understand the demands of the tasks in terms of perception, decision making and action. This would place requirements on the ML which would need to be demonstrated.

Number	Principle	Comment
EHF6	Workspaces in which operations (including maintenance activities) are conducted should be designed to support reliable task performance. The design should take account of the physical and psychological characteristics of the intended users and the impact of environmental factors.	This principle may lead to the layout of workspaces, and signposts to assist in autonomous tasks. For example, signs, rails and barriers that are easy to detect. Diverse systems to assist with an autonomous robot understanding its location and situational awareness.
EHF7	Suitable and sufficient user interfaces should be provided at appropriate locations to provide effective monitoring and control of the facility in normal operations, faults and accident conditions.	See the previous discussion, as an example there may need to be layout changes or signs and sensors to assist an autonomous robot and reduce potential errors.
EHF8	A systematic approach to the identification and delivery of personnel competence should be applied.	Competence of the ML and autonomous systems would need to be demonstrated, we assume with V&V evidence. Currency and checking that long term performance is as expected would be needed.
EHF9	Procedures should be produced to support reliable human performance during activities that could impact on safety.	As an example, there may need to be monitors to assist or police an autonomous robot and reduce potential errors.
EHF11	There should be sufficient competent personnel available to operate the facility in all operational states.	Sufficient ML/autonomous systems would need to be provided, such as including redundant systems.
EHF12	A management process should be in place to ensure the fitness for duty of personnel to perform all safety actions identified in the safety case.	This applies but in the engineering sense. Procedures might be needed to ensure an autonomous system or ML was fit for a new task.
EHF10	Human reliability analysis should identify and analyse all human actions and administrative controls that are necessary for safety.	This applies but in the engineering sense. The failures of the ML should be understood, and their likelihood. This may require understanding of confidence measures and monitoring of performance.

Table 11: Applying the human factors SAPs to ML systems directly

Table 11 provides some speculative discussion of how human factors related SAPs might apply to autonomous systems. Whilst they have provided an interesting prompt for discussion points (for example changes to workspaces) it may be the case that similar engineering-based principles would be more appropriate if an actual analysis was performed.

3.6 Security

In Section 2.3.4 we discussed the importance of security and of security-informed safety. As we noted, security is of particular importance to AI/ML systems, where the attack surfaces are substantial and different to those of traditional software. Paragraphs 39-41 of the SAPs outline security requirements. Security is also discussed briefly in EKP.5 where it is noted that safety and security should be treated in a complementary manner and not compromise one another. Security requirements are largely deferred to the Security Assessment Principles (SyAPs).

The SyAPs provide a comprehensive set of fundamental security principles and guidance covering

- Leadership and Management for Security
- Organisational Culture
- Competence Management
- Nuclear Supply Chain Management
- Reliability, Resilience and Sustainability
- Physical Protection Systems
- Cyber Security & Information Assurance
- Workforce Trustworthiness
- Policing and Guarding
- Emergency Preparedness and Response

As with the SAPs, they provide high level principles that could inform and be interpreted for the particular challenges of AI/ML. Similarly, the importance of data arises as a cross-cutting theme to be addressed. As outlined in Section 2.3.4 there is generic guidance being published by CPNI that could be built upon for the nuclear industry. There are two converging issues: the need to address security-informed safety and the challenges of AI/ML.

3.7 Safety cases

A thorough understanding of the themes discussed above is required in order to build a strong safety case, however the construction of safety cases, in itself, raises additional considerations for AI/ML which are discussed in this section. Table 12 presents comments on relevant sections from the SAPs, which are discussed in more detail below.

Reference	Relevant guidance	Comments
SC.4	"A safety case should: (a) explicitly set out the argument for why risks are ALARP;"	How can AI/ML systems be shown to be ALARP?

Reference	Relevant guidance	Comments
SC.7	"The safety case will also need to be updated to take account of changes at the facility, the site and its surroundings, for instance: (a) changes arising from modifications or revised operating methods or processes;"	How do AI/ML systems that update or learn 'on the job' impact this?
SC.2	"It is essential that the safety case documentation is clear and logically structured so that the information is easily accessible to those who need to use it."	What is the best structure in order to document an AI/ML safety case?

Table 12: Guidance impacted by AI/ML systems in safety case development

Demonstrating an AI/ML system's risks are ALARP poses novel challenges. For traditional software, this is often demonstrated by adherence to standards or other accepted best practices, in their development process. Standards such as IEC 61508 are widely accepted, with established quality. Such standards and practices are less well developed in the AI/ML industry (see the discussion on standards in Appendix A). Moreover, identifying what hazard analyses need to be undertaken, and how risks have been mitigated, may prove more complex due to the black box nature of AI/ML systems.

The novelty and additional complexity of AI/ML systems compared to conventional systems inevitably introduce additional risks. To demonstrate that the risks of an AI/ML system are ALARP, it is therefore very likely that the safety case will need to demonstrate that the use of an AI/ML component is necessary. This is addressed by ESS.21: "The design of safety systems should avoid complexity, apply a fail-safe approach and incorporate means of revealing internal faults at the time of their occurrence." In particular, ESS.21 states that where complexity cannot be avoided, the safety case should contain a "comprehensive examination of all the relevant scientific and technical issues". In the context of an AI/ML system, this should include a justification for the choice of AI/ML techniques used, and the reasons why the same functionality could not be achieved in a simpler way without AI/ML.

AI/ML systems that update may cause the safety case to expire more rapidly. Whilst a safety case is already seen as a "living document", adding continuously learning AI components will significantly speed up its "heartbeat"; as such, it may need updating much more regularly. This includes the wider issue of working in a fast moving industry with rapidly evolving technology and best practice.

The problem of constantly adapting systems may be solved simply by not allowing these algorithms (they comprise only a subset of AI/ML systems). ESS.15 alludes to this: "No means should be provided, or be readily available, by which the configuration of a safety system, its operational logic or the associated data (trip levels etc.) can be altered, other than by specifically engineered and adequately secured maintenance/testing provisions used under strict administrative control.". If these types of AI/ML systems are to be considered by ONR for assuring, this guidance, and the guidance on safety case updating, need to be carefully considered.

Finally, clear and logical safety cases must be developed. AI/ML systems are often complex, non-transparent and with hidden behaviour. It is essential that the safety principles as applied to them can be expressed clearly and logically. In Section 4.2.3 we outline our research into assuring AI systems using the newly developed Assurance 2.0 approach.

While we have focused on the issues of assuring AI/ML-based systems, we should not ignore the potential of AI/ML and other technologies to support the synthesis and maintenance of safety cases. For example, Adelard is currently working as part of a research project within the DARPA programme on the automation of certification [56].

3.8 Data management

A final and significant theme, not discussed in detail in the SAPs, is data management for AI/ML.

Whilst data is discussed in the context of building computational models of plant systems, reliability data and other plant data, data usage for AI/ML products is qualitatively different. Unlike traditional software, AI/ML products rely on data for training, testing and ongoing learning. For these devices, data management is as impactful as software development, in that all the system's behaviour is based on the combination of the training data and the model used to process it. The claimed reliability of the system is typically dependent on the accuracy and relevance of the test data.

In Section 4.2.2 we outline some data management principles that can allow for safe deployment of these systems.

3.9 Discussion and summary

In this section, we have surveyed the existing regulatory guidance and highlighted key areas where attempts to assure an AI/ML product may bring about ambiguity or limitations. In general the SAPs as overarching principles are still applicable, however they require additional interpretation and guidance for this incoming paradigm shift. The influence of AI/ML on regulation is cross cutting, and so will have a widespread impact on the SAPs.

In Section 4, we will discuss how the additional guidance and clarity can be given such that ONR is able to best make use of this innovative and disruptive technology.

4 Route map towards supporting AI/ML assurance

In this section, we highlight elements of a work programme to underpin ONR's aim of ensuring that regulation does not present unnecessary barriers to the deployment of AI/ML systems, and that they have the capability to adequately assess systems using this technology. This involves recommendations to both updating the SAPs (as investigated in Section 3), and building an overall AI strategy. Specifically, we recommend that ONR considers four areas of focus:

- **Developing an AI regulatory framework, building upon SAPs and guidance**
    - Clarify the role and types of technologies used in AI/ML systems via taxonomies and automation levels.
    - Consider interpretations and changes to the SAPs (discussed in Section 3) to address human factors, PE/ICBMs, security, understanding and data issues. This may require nuclear industry specific research to address.
    - Consider more fine-grained use of claims, arguments, evidence as well as property-based approaches to assurance to replace PE/ICBMs.
- **Taking an active role to build capability within industry and ONR**

  - By taking an active lead in research, trials and benchmarking, ONR can help build systems and safety cases that fit the needs of the UK nuclear sector.
  - Research streams include data, human factors and sociotechnical systems, computer architectures, hazard analysis, security-informed safety, and confidence building technologies.

- **Developing architectural approaches and a data strategy**
  - Develop additions to the SAPs to address the role of data and its evaluation.
  - Identification and development of analysis techniques for assessing data properties and provenance.
  - Research the role of architecture and data and their impact on safety justification and risk e.g., defence in depth, diversity, and risk control hierarchy.
- **Engaging with standards**
  - Engage in a focused manner with the standardisation process.

These are first summarised in terms of an initial route map and then key elements are elaborated in the following subsections. Implicit in the strategy is the need to develop an education and training plan to ensure a pipeline of expertise is maintained.

In Table 13 we summarise the initial route map in terms of three overlapping phases:

- Phase 1 – 1-2 years
- Phase 2 – 1-5 years
- Phase 3 – 3-10 years

In the longer term we can imagine justifying the coworking of humans and AI/ML systems and a general increase in scale and criticality and decrease in cost. We have also identified a number of potential blockers to the strategy. Of course, these time scales are subject to change depending on the speed of development and the appearance of potential blockers.

	Phase 1	Phase 2	Phase 3
Regulator capability and framework	Research and development of an AI/ML regulatory framework to update SAPs and guidance, including rationale. Explore regulatory approaches to risk control that take into account trade-offs between benefits and risks, in the face of uncertainty.	Update the SAPs and supporting guidance. The first systems with AI/ML components accepted for low integrity applications.	Critical and co-operative systems justifiable e.g., AI/ML-based systems in control rooms. Use of generalised AI tools justifiable.
Capability (analysis techniques and evidence generation)	Research and development of evidence generation and analysis techniques for safety cases, and safety case methodologies. Establish data collection techniques to make best use of operating experience to inform safety cases.	Techniques for assuring algorithm supply chains are established. Deployment of evidence generation and analysis techniques on real examples. Approaches to assess algorithms in silicon established.	Scale increased, criticality increased, cost reduced.
Data and architectural approaches	Data route map established. Research initiated on auditability of data and priorities for nuclear industry. Research impact of architecture and system analysis for AI/ML systems and associated data e.g., role of monitor architectures, defence in depth and diversity. Understand the role of the hierarchy of risk controls in an AI/ML contexts.	Techniques for assessing data properties, supply chains and safety developed and trialled in nuclear context. System architectures and associated analysis techniques established.	Scale increased, criticality increased, cost reduced. More reuse and revalidation of data.
Standards	Standards route map defined. Routes for validation of standards addressed. Gaps in standards addressed, premature standards discouraged.	Role and utility of emerging standards assessed. Cross-sector experience captured and evaluated. Guidance developed on standards interpretation and use.	Standards validation and impact assessed. Standards for generalised platforms initiated.
Potential blockers	Slow pace of change in regulation. Real world compromises of embedded AI/ML systems. Proliferation of poor standards and poor practice. Difficulties in initiating a long term well-resourced strategy.	AI/ML research not focused on embedded system safety. Incidents (e.g., involving AVs) eroding confidence in AI/ML. Resource and expert scarcity. Threat actors making increased use of AI/ML. Lack of licensee confidence and competence.	Low public acceptance of AI/ML. Complexity of legal and liability issues for licensees.

Table 13: Summary route map

4.1 Developing an AI framework

In contrast to traditional computer systems, AI/ML systems may be required to make autonomous decisions (with or without supervision), or simply provide advice to a human operator. It is clear that the level of autonomy given to a system greatly impacts the appropriate level of trustworthiness required. This adds an additional dimension to the level of safety assessment required, where both the criticality of the safety function and the level of autonomy decide what safety precautions are needed. This is illustrated, with examples, in Table 14. The more critical the safety function, and the more autonomous the system, the more trustworthiness is needed.

The safety criticality dimension describes the overall contribution of the ML component to the safety role. It might be related to the safety classification of the function or categorisation of the equipment, the risk based role of the system, or some combination of controllability and consequence. Other approaches might seek to combine the safety benefit of undertaking the project with the safety risks of not doing so.

Alternative frameworks for autonomy have been suggested, for example those which divide systems into Reactive (with a strict envelope of operation), Rules-based (where a combination of a model of the environment and regulations dictate behaviour) and Principle-based (which are able to perform even in unexpected situations) [65].

		Level of Autonomy		
		System provides operator assistance (advice)	Conditional automation (advice and supervised action)	Full automation for part of mission
Safety criticality	Low	Video feed augmented with personnel data.		Exploratory rover.
	Medium	Alerting possible off-normal behaviour detected in process vessels, as part of international safeguarding.	Monitoring multiple video feeds, selecting which appears on the main monitor for an operator.	Autonomous crane system with automatic object detection.
	High	Alarm system detecting possible faults in the core.	Automatous control system with key decisions requiring authorisation from a human operator.	Fully autonomous protection system.

Table 14: Example AI technologies for various levels of autonomy and safety class

Whilst the precise framework that would be most beneficial requires further analysis, we recommend that a detailed discussion of these levels is needed in some form in the SAPs or associated guidance to enable discussion and common view of the assurance challenge.

The safety criticality and level of autonomy could be combined into a single index that is used to define a level of "assurance challenge". Table 15 provides an illustration of what this could look like, where the cells represent a graded approach. AC4 is the highest level of challenge.

		Level of Autonomy of the AI/ML role		
		System provides operator assistance (advice)	Conditional automation (advice and supervised action)	Full automation for part of mission
Safety criticality	Low	ACL1	ACL2	ACL3
	Medium	ACL2	ACL3	ACL4
	High	ACL3	ACL4	ACL4

Table 15: Assurance challenge levels (ACLs) – illustrative only

In addition, these tables together with taxonomies on types of AI/ML and their assurance challenges could inform a progressive deployment strategy. Here, a phased approach is used to assure simpler, less safety-critical systems first, and use lessons learnt and gained experience to progressively work with more complex and safety-critical systems over time. It could be used to shape industry expectations and profile ONR capabilities.

The AI systems most easily integrated in the short term lie in the left of Table 14; these simply provide advice to a human operator, without any autonomy. Such devices are already deployed in a range of sectors, such as security and medicine, and could have a wide range of applications within a nuclear power plant. Close monitoring and engagement with these sectors to learn how these industries are working to assure these technologies is essential.

4.2 Building upon regulatory principles and guidance

There are a wide range of topics that need to be addressed covering human factors, PE/ICBMs, security, understanding and especially data issues. Here we elaborate on three key issues: understanding, safety assurance and data.

4.2.1 Clarifying the discussion of "understanding" can support assurance

As highlighted in Section 3.2, the extent to which duty holders are required to understand the behaviour of AI/ML systems requires clarification in the SAPs. AI raises additional nuance in the discussion of understanding, as its underlying processes are often uninterpretable even to experts with a detailed knowledge of the device. The SAPs therefore need to better define the role of understanding, and give additional information on what is expected of AI/ML systems.

We anticipate certain levels of knowledge abstraction are acceptable – for example, in most circumstances, fully explainable AI is likely to not be required. Note that certain amounts of knowledge abstraction are already in place in safety cases; for example, complex mathematical models to describe processes such as probabilistic fracture mechanics are not required to be understood in detail in order for the stakeholders to claim an understanding of the safety case. As such we expect stakeholders are only required to have an analogous "reasonable understanding" of the AI/ML systems present in the safety case, which should depend on the level of autonomy given to it.

In Table 16 we outline three possible levels of understanding. For an AI/ML product to be assured, the identified level of understanding should be justified by its relative level of autonomy and safety criticality. These have been chosen so as to give clear and verifiable requirements for understanding.

In order to demonstrate competence and responsibility for the AI/ML products in safety-critical systems, stakeholders should:	
Level 1	• Understand the probabilistic reliability of the product in the application scenario, and the metrics used to measure this. • Understand how the ML method is implemented and applied. o Example: The ML component uses a deep neural net, due to the high-dimensional features of the input data. • Understand the extent to which the AI/ML product updates its learning whilst operational. Static products are trained once, whilst dynamic products continually update their knowledge. o Example: This ML sensor receives operator feedback when a false positive alert is sent. Therefore operators must be aware of this and trained appropriately.
Level 2	• Understand in detail the scope of the training data, and therefore what circumstances the ML method may be extrapolating beyond its training and therefore have poorly understood behaviour. o Note: This is referenced in ESS.10 – "The capability of a safety system…should be defined". o Example: The autonomous rover has not been trained in snow, therefore its deployment under these conditions would be an extrapolation from its training data and therefore dangerous. • Understand at a high level the reasons decisions were taken. o Example: This route was chosen based on the fastest route. Alternative routes are described in the log file.
Level 3	• Be able to walkthrough past and hypothetical decisions in detail. • Be given human understandable 'reasons' for making decisions.

Table 16: Three levels of understanding an AI/ML system

Whilst Levels 1 & 2 are feasible with current technology, Level 3 is likely beyond the capabilities of the majority of today's commercial products.

Note that some researchers consider it unacceptable to accept any safety-critical decision of an AI product working in a safety-critical setting without it being able to explain why it made that decision [66]. As such, it may be judged that a fully autonomous AI is not at a level of maturity appropriate for the most safety-critical systems.

4.2.2 Guidance on data is needed

As discussed in Section 3.8, there is a regulatory gap for data management for AI/ML systems. Data management corresponds to the quality assurance of data used to train and test the system, data that is processed as the system operates and control of dataset drift. Guidelines for data management in AI have been widely suggested [67], so the inclusion of data management principles or guidelines in the SAPs is both feasible and essential.

Some key data management principles for AI/ML are

- The data used to train an ML algorithm must be both accurate and representative for the intended application.
- When selecting or creating training datasets, it must be the case that diversity, historic bias, ethics, privacy and fairness are all explicitly considered.
- Edge cases and rare events should be enhanced in the training data and the test cases.
- It must be ensured personal data is held securely, and cannot be used to discriminate against persons. Some types of big data analytics, such as profiling, can have intrusive effects on individuals, and the complexity of ML methods can make transparency difficult.
- Training and test data must be controlled carefully. If the test data is contaminated with information from the training data, this can lead to over-optimistic results.
- Novel attack surfaces, through the poisoning or acquisition of the training data, must be accounted for.
- AI/ML systems should be auditable, with an appropriate data log.

These data management principles must be clarified in the SAPs in order to ensure safe AI/ML systems. In addition there need to be corresponding tools and methods for assessing the data properties, the architectural mitigations to data issues and supply chain security.

4.2.3 Safety and security case guidance

Safety and security cases may require different structure and argumentation in order to be appropriate for AI/ML assurance. Issues such as ALARP arguments, safety case refreshing and novel argumentation were discussed in Section 3.7. In our work in TIGARS [10] we argued that assuring trust and trustworthiness through argument-based mechanisms, specifically CAE, allowed for the accelerated exploration of novel mechanisms (e.g., architectural approaches) that could lead to advancements in the assurance of disruptive technologies.

In [10] we summarise the conclusions and overall recommendations of this work. In terms of assurance we concluded

1. Developing an assurance strategy should be a key part of the overall design approach and integrated into the overall lifecycle. The assurance approach should be commensurate with the different risks and be consistent across them, e.g., by adopting an outcome-based, risk-informed approach.
   1.1. Novel assurance approaches (e.g., articulated using CAE) exclusive to ML and AI-based systems should be developed to identify areas to focus on and establish how they impact both the system and its assurance. It can help define and evaluate the reasoning and evidence needed.
   1.2. Key claims should address the high-level functional and ethical principles such as those from the EU Expert Group report [12] and the SHERPA project [35]. These principles can be used to shape and define system or service level properties.
   1.3. An assurance case for autonomous systems should at a minimum address the points below:
      – what the system is and in what environment and ecosystem it will operate in
      – how much trust in a system is needed, considering interdependencies and systemic risks
      – whether it is sufficiently trustworthy to be initially deployed

   - whether it will continue to be trustworthy in the face of environmental changes, threat evolution and failures

2. Structured argumentation for safety cases (and more generally assurance cases) needs more emphasis on reasoning and evidence, if the cases are to be sufficiently robust and acceptable. We have characterised a new CAE-based assurance framework to achieve this, which would utilise evidence extracted from V&V, defence in depth, and diversity techniques.

The last point is based on the observation that the CAE or GSN structure for an AI/ML case might be similar to that of a conventional technology's, but they key underlying difficulties are evidence and reasoning (e.g., how to assess the performance of an ML-based sensor). Subsequent work has led to "Assurance 2.0: A Manifesto" [36] aiming to enable the innovation and continuous incremental assurance, perhaps counterintuitively, by making assurance more rigorous. This increased rigour comes from an increased focus on the reasoning and evidence employed as well as an explicit identification of defeaters and counterevidence. Key elements of Assurance 2.0 are

- Making explicit inference rules and the separation of inductive and deductive reasoning.
- The use of an Indefeasibility Criterion for justified belief to frame the use of defeaters - both undercutting and rebuttal.
- Focus on evidence integration, addressing both the relevance and provenance of evidence.
- Confirmation theory to evaluate the strengthening of evidence and arguments: it is not enough for evidence to support a claim; it must also discriminate between a claim and its negation or counterclaim.
- Explicit approach to reduce bias by the use of counter-cases and the aforementioned confirmation theory.
- Recognition of importance of both mindset and methodology.

Note that Assurance 2.0 only provides the framework or set of concepts for developing cases. It needs to be supported by specific approaches to reasoning and evidence. To support these, there is a need for specific research on reasoning about confidence in autonomous operation. Some of these use a variety of Bayesian frameworks which build on the work on machine learning and also earlier research on software reliability modelling. For example, in [37], the authors present a new variant of Conservative Bayesian Inference (CBI), which uses prior knowledge while avoiding optimistic biases. CBI is used to assess the reliability of autonomous vehicles by applying Software Reliability Growth Models (SRGMs) to operating experience, specifically disengagement data (take-overs by human drivers). Related work [38] extends this by proposing a property-based decomposition of the safety case and assessing the system in two steps. The first step is based on assurance activities conducted at each stage of its lifecycle, e.g., formal verification on the neural network robustness. The second step boosts the confidence using field data of successful operation and a CBI approach. CBI is discussed in more detail in Appendix B.

Addressing the security-informed safety aspects could consider building on the CPNI guidance discussed in Section 3.6.

4.3 Engaging with standards

Standards are an important part of regulation in defining accepted good practice and hence facilitating judgments of proportionality. In addition, in dealing with computer-based systems, they are key in defining the excellence of production approach.

As discussed in Section 2.3.3, there is currently a wide range of standardisation and guidance activities relevant to AI/ML, but many of the standards will lack the maturity of those associated with traditional software and hardware. Also, with so many activities it is likely that many of the documents will be derivative and/or of lower quality. An important aspect of understanding and applying these standards in the future is to monitor what experience is gained in other industries and sectors with the use of these standards.

It is recommended that ONR

- Engages with IAEA and IEC groups to ensure the standards groups are well informed and to influence the capturing of rationale for the standards.
- Monitors industries where AI/ML application is more onerous and also more advanced. For example, monitoring how UL 4600 is used in the automotive industry.
- Undertakes occasional scanning of the standardisation and guidance landscape to identify useful input from outside of the industry (e.g. in business technology, in security or defence).

4.4 Taking an active role in research

There are a number of research initiatives that ONR could monitor and shape to maximise benefits for the nuclear industry. These include those specific to the nuclear sector such as RAIN (Robotics and AI in Nuclear) but also the newly announced UKRI TAS Node on Governance and Regulation. In this section we outline two possible research avenues that would be beneficial to ONR.

4.4.1 Architectural approaches

Architectural approaches to building AI/ML systems can provide a basis for reducing the assurance burden on the AI/ML component. This is particularly powerful for systems with safe states (or minimal risk states) that the overall system can move to when stress is detected, and is discussed in detail in Appendix C. Since these are built from traditional software components, this can greatly simplify the assurance process.

To design and deploy a safety monitor one needs

- Hazard analysis techniques that allow the behaviour of the AI/ML component and the overall monitored system to be assessed.
- Understanding and justification of the metrics and signals that can be used to monitor and benchmark the performance of the system. A discussion of the need for metrics and benchmarks is provided in Appendix B.
- A graded approach to describing the role of the AI/ML component (so this might challenge the industry classification and categorisation approaches).
- A graded approach to assuring AI/ML systems.

These topics are the subject of ongoing research and present a promising route into deploying AI/ML systems in power plants.

4.4.2 Analysis techniques and evidence generation

AI/ML systems require new analysis and evidence generation techniques. These include hazard analysis, statistical testing (including the use of varieties of environment and platform simulations), static analysis and the role of metrics and benchmarks.

As such, novel templates for safety cases and associated reasoning techniques need to be developed. Our work developing safety case templates for autonomous systems in [39] produced several template blocks for strengthening safety cases. These 'CAE template blocks' covered a range of assurance issues, within the UK regulatory context, such as requirements specifications, hazard analyses and sensors.

Novel evidence generation techniques also require research. One cause of this is the large number of trials required to build confidence in an AI/ML system due to its black box nature and high-dimensional input space. An example of a type of argument that can support these claims is Conservative Bayesian Inference. This allows for systematic and quantitative incorporation of test data with previous experience and lifecycle information. This is discussed in detail in Appendix B.

Further research into developing these reasoning methods and CAE template blocks is essential to building strong safety cases for AI/ML systems.

5 Summary and conclusions

The overall aim of this project is to advise ONR on the suitability of existing UK nuclear regulation with regards to the application and use of Machine Learning (ML) and Artificial Intelligence (AI) in operations affecting nuclear material. This report draws on specific research undertaken for ONR and other research being undertaken by Adelard on assuring autonomous systems.

We provide background and an overview of the AI/ML landscape, focusing on the types of systems available, nuclear-specific applications and a discussion of the difficulties in regulating and assuring them. AI/ML systems can be viewed on a spectrum from a software-centric view, to a more anthropomorphic viewpoint, where we treat AI/ML systems as operators with levels of autonomy and authority.

We have reviewed ONR's safety assessment principles to identify areas that may be effected by, or support, AI/ML assurance. Whilst the safety assessment principles themselves remain strong, there is need for clarification and additional guidance. These updates could take the form of augmenting the SAPs across the document, creating separate principles that account for, and clarify, AI/ML specific topics, or through the inclusion of an additional technical assessment guide to cover topics such as data, security and autonomy frameworks.

We highlighted elements of a work programme – a high-level route map – to underpin ONR's aim of supporting UK nuclear facilities to take advantage of AI/ML. This involves recommendations to both updating the SAPs and building an overall AI strategy. Specifically, we recommend that ONR consider four areas of focus:

- **Developing an AI regulatory framework, building upon SAPs and guidance**
    - Clarify the role and types of technologies used in AI/ML systems via taxonomies and automation levels.
    - Consider interpretations and changes to the SAPs (discussed in Section 3) to address human factors, PE/ICBMs, security, understanding and data issues. This may require nuclear industry specific research to address.
    - Consider more fine-grained use of claims, arguments, evidence as well as property-based approaches to assurance to replace PE/ICBMs.
- **Taking an active role to build capability within industry and ONR**

  - By taking an active lead in research, trials and benchmarking, ONR can help build systems and safety cases that fit the needs of the UK nuclear sector.
  - Research streams include data, human factors and sociotechnical systems, computer architectures, hazard analysis, security-informed safety, and confidence building technologies.
- **Developing architectural approaches and a data strategy**
  - Develop additions to the SAPs to address the role of data and its evaluation.
  - Identification and development of analysis techniques for assessing data properties and provenance.
  - Research the role of architecture and data and their impact on safety justification and risk e.g., defence in depth, diversity, and risk control hierarchy.
- **Engaging with standards**
  - Engage in a focused manner with the standardisation process.

AI/ML is likely to become a very varied and pervasive technology (just as software has become). AI/ML is both an opportunity and challenge to ONR and, as this work shows, will touch many safety areas that ONR regulate. This work highlights the breadth and tempo of the strategic response needed from ONR: the need to innovate in interpreting and developing the SAPs, the multi-disciplinary nature of the task, and the wider challenge of data and analysis technologies to support assurance of systems with AI/ML components.

6 Glossary

AAIP	Assuring Autonomy International Program
AI	Artificial Intelligence
ALARP	As Low As Reasonably Practicable
AWI	Accepted Work Item
ASP	Answer Set Programming
CAE	Claims, Arguments, Evidence
CBI	Conservative Bayesian Inference
CD	Committee Drafts
CPNI	Centre for the Protection of National Infrastructure
DARPA	Defence Advanced Research Projects Agency
HIC	Human-in-command

HITL	Human-in-the-loop
HOTL	Human-on-the-loop
HFI	Human Factors Integration
HMI	Human Machine Interface
HRA	Human Reliability Analysis
IAEA	International Atomic Energy Agency
ICBMs	Independent Confidence Building Measures
LEC	Learning Enabled Component
LSAD	Low-Speed Automated Driving
ML	Machine Learning
OSD	Open Systems Dependability
PE	Production Excellence
PP	Predictive Processing
RAS	Real-time Autonomous System
SAP	Safety Assessment Principle
SASWG	Safety of Autonomous Systems Working Group
SCSC	Safety-Critical Systems Club
SOTIF	Safety Of The Intended Functionality
SME	Subject Matter Expert
SQEP	Suitably Qualified and Experienced Person
SRGM	Software Reliability Growth Model
SyAPs	Security Assessment Principles
TAG	Technical Assessment Guide
TIGARS	Towards Identifying and closing Gaps in Assurance of autonomous Road vehicleS
TR	Technical Report
TTN	TIGARS Topic Notes
UAV	Unmanned Aerial Vehicles
V&V	Verification and Validation

XAI	Explainable AI

7 Acknowledgements

The work here benefits from previous and current Adelard work. In particular, the TIGARS project TIGARS Topic Notes reported in [10] and the work in [39] on Safety Case Templates and Guidance. The Tigars Topic Notes will be published by CPNI this spring.

8 Bibliography

[1] R. Haddal, N. Hayden, S. Frazar, *Autonomous Systems, Artificial Intelligence and Safeguards*, IAEA Symposium on International Nuclear Safeguards: Building Future Safeguards Capabilities held November 5-8, 2018, Vienna, Austria.

[2] Dharmendra S. Modha, Introducing a Brain-inspired Computer, TrueNorth's neurons to revolutionize system architecture, See https://www.research.ibm.com/articles/brain-chip.shtml

[3] Beyond Today's AI, New algorithmic approaches emulate the human brain's interactions with the world, https://www.intel.co.uk/content/www/uk/en/research/neuromorphic-computing.html accessed 16/10/20

[4] RAIN, https://rainhub.org.uk/. Accessed October 2020.

[5] IAEA report, https://www.osti.gov/servlets/purl/1561151

[6] UL 4600, Standard for Safety for the Evaluation of Autonomous Vehicles and Other Products. Underwriters Laboratories, https://ul.org/UL4600. Accessed September 2020.

[7] PAS 1881, Assuring safety for automated vehicle trials and testing – Specification

[8] M Holloway, Understanding the Overarching Properties, NASA/TM–2019–220292, July 2019

[9] Erin E. Alves et al, Considerations in Assuring Safety of Increasingly Autonomous Systems, NASA/CR–2018-220080

[10] TIGARS, Overview and Introduction to The TIGARS Topic Notes, D5.6: D/1259/138008/7 v1.0, 2020. https://arxiv.org/pdf/2003.00789.pdf.

[11] Robustness Testing of Autonomy Software, C Hutchison et al. IEEE/ACM 40th International Conference on Software Engineering: Software Engineering in Practice Track (ICSE SEIP), May/June 2018.

[12] Ethics Guidelines for Trustworthy AI, the High-Level Expert Group on Artificial Intelligence (by the European Commission). 8 April, 2019.

[13] SASWG, Safety assurance objectives for autonomous systems, V2.0 SCSC-153A, 978-1654029050, 2020.

[14] Brundage, Miles, Shahar Avin, Jian-Bing Wang, Haydn Belfield, G. Krüger, Gillian K. Hadfield, Heidy Khlaaf, et al. "Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims." ArXiv abs/2004.07213 (2020).

[15] ONR, Security Assessment Principles for the Civil Nuclear Industry, Version 0, 2017.

[16] Bloomfield R., Butler E., Guerra S., and Netkachova K., Security-informed Safety: Integrating Security within the Safety Demonstration of a Smart Device, Nuclear Plant Instrumentation, Control, 11-15 June 2017.

[17] DHS (2011) DHS evidence "Hearing Before The Subcommittee On National Security, Homeland Defense And Foreign Operations Of The Committee On Oversight And Government Reform House Of Representatives One Hundred Twelfth Congress First Session, May 25, 2011, Serial No. 112–55".

[18] Cyber Security at Civil Nuclear Facilities: Understanding the Risks, C. Babylon, R. Brunt, and D. Livingstone, Chatham House, Royal Inst. of Int'l A airs, 2015.

[19] Confirmation of a Co-ordinated Attack on the Ukrainian Power Grid, Assante (2016), M.J. Assante, blog, 9 Jan. 2016; https://ics.sans.org/blog/2016/01/09/confirmation-of-a-coordinated-attack-on-the-ukrainian-power-grid. Accessed 12 November 2016.

[20] TIGARS Topic Note, Security-informed safety analysis, D5.6.7 (W3022). January 2020.

[21] Security-Informed Safety: If it's not secure, it's not safe, Bloomfield (2013), R. E., Netkachova, K. & Stroud, R. Software Eng. for Resilient Systems, A. Gorbenko, A. Romanovsky, and V. Kharchenko, eds., LNCS 8166, Springer, 2013, pp. 17–32.

[22] The risk assessment of ERTMS-based railway systems from a cyber security perspective: Methodology and lessons learned, Bloomfield, R. E., Bendele, M., Bishop, P. G., Stroud, R. & Tonks, S. (2016). Paper presented at the First International Conference, RSSRail 2016, 28-30 Jun 2016, Paris, France.

[23] IEC 61882:2016 Hazard and operability studies (HAZOP studies) - Application guide.

[24] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199, 2013.

[25] Goodfellow, I., Shlens, J., Szegedy, C. "Explaining and harnessing adversarial examples," in International Conference on Learning Representations. Computational and Biological Learning Society, 2015.

[26] Benoît Bonnet, Teddy Furon, Patrick Bas. What if Adversarial Samples were Digital Images?. IH&MMSEC 2020 - 8th ACM Workshop on Information Hiding and Multimedia Security, Jun 2020, Denver, France. pp.1-11, ff10.1145/3369412.3395062ff. ffhal-02553006v2f

[27] K. Xu et al., Adversarial T-shirt! Evading Person Detectors in A Physical World, arXiv:1910.11099, July 2020.

[28] Nar K., Ocal O., Shankar S., Ramchandran K. Cross-Entropy Loss and Low-Rank Features Have Responsibility for Adversarial Examples, arXiv:1901.08360v1, Jan. 2019.

[29] Shamir A., Sefran I., Ronen E., Dunkelman O., A Simple Explanation for the Existence of Adversarial Examples with Small Hamming Distance, arXiv:1901.10861v1 Jan 2019.

[30] ONR Guide, *Computer Based Safety Systems*. Nuclear Safety Technical Assessment Guide NS-TAST-GD-046, Revision 5, April 2019.

[31] Dependability Assessment of Software for Safety Instrumentation and Control Systems at Nuclear Power Plants" (NP-T-3.27), https://www-pub.iaea.org/books/IAEABooks/12232/Dependability-Assessment-of-Software-for-Safety-Instrumentation-and-Control-Systems-at-Nuclear-Power-Plants last accessed March 2019.

[32] B. Littlewood, R.E. Bloomfield, The use of computers in safety-critical applications. London, UK: Health and Safety Commission, 1998.

[33] Lisanne Bainbridge, Ironies of automation, Automatica, Volume 19, Issue 6, 1983, Pages 775-779, ISSN 0005-1098 and also https://en.wikipedia.org/wiki/Ironies_of_Automation

[34] R.E. Bloomfield, W.D. Ehrenberger, Validation and licensing of intelligent software (IAEA-CN-49/68), Proceedings of An International Conference on Man-Machine Interface in The Nuclear Industry (Control And Instrumentation, Robotics and Artificial Intelligence), 1988.

[35] Shaping the ethical dimensions of smart information systems – a European perspective (SHERPA), Deliverable No. 3.2. BSI ART/1_19_025

[36] R. Bloomfield, J. Rushby, Assurance 2.0: A Manifesto. June 2020, https://arxiv.org/abs/2004.10474.

[37] Zhao, X., Robu, V., Flynn, D., Salako, K. and Strigini, L. (2019). Assessing the Safety and Reliability of Autonomous Vehicles from Road Testing. Paper presented at the ISSRE 2019 - the 20th International Symposium on Software Reliability Engineering, 28 - 31 October 2019, Berlin, Germany.

[38] Zhao, X., Banks, A., Sharp, J., Robu, V., Flynn, D., Fisher, M., & Huang, X. (2020). A Safety Framework for Critical Systems Utilising Deep Neural Networks. In A. Casimiro, P. Ferreira, F. Ortmeier, & F. Bitsch (Eds.), Computer Safety, Reliability, and Security. SAFECOMP 2020 (pp. 244-259). (Lecture Notes in Computer Science; Vol. 12234). Springer.

[39] R. Bloomfield, G. Fletcher, H. Khlaaf, L. Hinde, P. Ryan, Safety Case Templates for Autonomous Systems: Interim Report, Adelard, D1294v1.0, November 2020.

[40] Approved IEEE 7000™ Standards & Projects, IEEE Ethics in Action in Autonomous and Intelligent Systems, https://ethicsinaction.ieee.org/p7000/. Accessed September 2020.

[41] Standards Latest News, IEEE Robotics and Automation Society, https://www.ieee-ras.org/about-ras/latest-news/206-industry-government/standard. Accessed September 2020.

[42] StandICT.eu project. ICT standards and ongoing work at international level in the field of Artificial Intelligence (AI), https://www.standict.eu/artificial-intelligence-report last accessed in November 2019.

[43] ISO/IEC TR 24028:2020 - Artificial intelligence — Overview of trustworthiness in artificial intelligence.

[44] K. Shahriari and M. Shahriari, "IEEE standard review — Ethically aligned design: A vision for prioritizing human wellbeing with artificial intelligence and autonomous systems," 2017 IEEE Canada International Humanitarian Technology Conference (IHTC), Toronto, ON, 2017.

[45] IEEE 7010:2020 - Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being, 2020.

[46] H. Rezatofighi et al., Generalized Intersection over Union: A Metric and A Loss for Bounding Box Regression, Conference on Computer Vision and Pattern Recognition, 2019.

[47] O. Russakovsky, L.-J. Li, and L. Fei-Fei, Best of both worlds: human-machine collaboration for object annotation, Conference on Computer Vision and Pattern Recognition, 2015.

[48] G. Ciaparrone et al., Deep Learning in Video Multi-Object Tracking: A Survey, Neurocomputing Vol. 381, 2020.

[49] Thomas Metzinger & Wanja Wiese (Eds.), Philosophy and Predictive Processing (PPP), Frankfurt am Main: MIND Group, ISBN: 978-3-95857-138-9

[50] K. Basu, F. Shakerin, and G. Gupta, AQuA: ASP-Based Visual Question Answering, Practical Aspects of Declarative Languages, 2020.

[51] F3269-17 Standard Practice for Methods to Safely Bound Flight Behavior of Unmanned Aircraft Systems Containing Complex Functions, ASTM International.

[52] M Clark, X Koutsoukos, J Porter, R Kumar, G Pappas, O Sokolsky, I Lee, L Pike, A Study on Run Time Assurance for Complex Cyber Physical, AFRL/RQQA, 2013.

[53] Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. On a formal model of safe and scalable self-driving cars. ArXiv preprint arXiv:1708.06374, 2017.

[54] John Rushby. Runtime certification. In Martin Leucker, editor, Eighth Work- shop on Runtime Verification: RV08, volume 5289 of Lecture Notes in Computer Science, pages 21–35, Budapest, Hungary, April 2008. Springer-Verlag.

[55] Rafal Bogacz, A tutorial on the free-energy framework for modelling perception and learning, Journal of Mathematical Psychology, Vol 76, Part B, Feb 2017: doi.org/10.1016/j.jmp.2015.11.003

[56] See Automated Rapid Certification Of Software (ARCOS), https://www.darpa.mil/program/automated-rapid-certification-of-software

Appendix A
Standards and guidelines landscape

In Section A.1, we outline both developing and completed standards of interest that are guiding the standardisation roadmap of AI and ML. We then provide a more in-depth overview of candidate standards which are available and their applicability to the nuclear domain.

A.1 Landscape review

In this section, we expand on our literature review of standards which have appeared since the Towards Identifying and closing Gaps in Assurance of autonomous Road vehicleS (TIGARS) project, as part of the Assuring Autonomy International Program (AAIP), in which a review of international standards for autonomous systems was performed.

The TIGARS report [10] provided an overview of standards and guidance relevant to assurance of Real-time Autonomous Systems (RASs). Standards, guidance, and white papers were selected and reviewed using the following: reference to the source material, summary of the document, relevance to assurance of RAS and observations and recommendations.

The report focuses on several areas of assurance challenges and the standards being developed in each. Table 17 shows the relevant areas considered in the report and the mapping to standards and policy papers.

Assurance area	Standards
Requirements and testing for RAS	ISO 22737 is a standard for requirements and testing of low-speed automated driving (LSAD) systems. ISO/PAS 21448 is a standard on safety of the intended functionality (SOTIF) for autonomous road vehicles.
Safety assurance of RAS	Uber Advanced Technologies Group. A Principled Approach to Safety. 2018. FiveAI. Certification of Highly Automated Vehicles for Use on UK Roads -- Creating an Industry-Wide Framework.
Guidance on AI	OECD (Organisation for Economic Co-operation and Development). The Recommendation on Artificial Intelligence – the first intergovernmental standard on AI. UNESCO and COMEST. Preliminary study on the ethics of artificial intelligence. SHS/COMEST/EXTWG-ETHICS-AI/2019/1. IEC White Paper: Artificial intelligence across industries. Google "Responsible AI Practices".

Table 17: Summary of TIGARS review

In the remainder of this section, we further outline both developing and completed standards of interest that are guiding the standardisation roadmap of AI and ML. Specifically, we focus on standards that would be relevant or applicable to safety-critical systems or the nuclear domain.

A.1.1 ISO/IEC AI standards

The technical committee ISO/IEC JTC 1/SC 42 — Artificial Intelligence was established in October 2017 and focuses on developing AI applications. As this group was only recently established, SC 42 has not yet published many International Standards, with the exception of ISO/IEC TR 24028:2020 *Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence*. There are currently other Accepted Work Items (AWIs) and Committee Drafts (CD). We give an overview of such relevant standards below.

Reference	Title	Status
ISO/IEC 22989	Artificial intelligence — Concepts and terminology	CD
ISO/IEC 23053	Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)	CD
ISO/IEC TR 24372	Information technology — Artificial intelligence (AI) — Overview of computational approaches for AI systems	AWI
ISO/IEC 38507	Information technology — Governance of IT — Governance implications of the use of artificial intelligence by organizations	CD
ISO/IEC 24668	Information technology — Artificial intelligence —Process management framework for Big data analytics	AWI
ISO/IEC 23894	Information Technology — Artificial Intelligence — Risk Management	CD
ISO/IEC TR 24372	Information technology — Artificial intelligence (AI) — Overview of computational approaches for AI systems	AWI
ISO/IEC TR 24368	Information technology — Artificial intelligence — Overview of ethical and societal concerns	AWI
ISO/IEC TR 24029-1	Artificial Intelligence (AI) — Assessment of the robustness of neural networks — Part 1: Overview	CD
ISO/IEC TR 24027	Information technology — Artificial Intelligence (AI) — Bias in AI systems and AI aided decision making	AWI

Table 18: ISO/IEC SC 42 relevant AWI and CD standards

The release schedule for the above standards is not known or available. The published ISO/IEC 24028 surveys topics related to trustworthiness in AI systems, including

- approaches to establish trust in AI systems through transparency, explainability, controllability, etc.
- engineering pitfalls and typical associated threats and risks to AI systems, along with possible mitigation techniques and methods
- approaches to assess and achieve availability, resiliency, reliability, accuracy, safety, security, privacy, maintainability, and durability of AI systems

Separately, the Software and Systems Engineering Technical Committee ISO/IEC JTC 1/SC 7 is developing ISO/IEC/TR 29119-11 *Software and systems engineering — Software testing — Part 11: Testing of AI-based systems*, to further address the verification of AI systems.

A.1.2 IEEE Standards Association

The IEEE Standards Association is pursuing several threads in the domain of AI, aiming to target both general and domain specific applications of AI. The IEEE P7000 series of standards under development emphasise the importance of certified accountability, transparency and reduction of algorithmic bias as being a critical enabler for AI realisation. Whereas more traditional standards have a focus on integration and safety, the IEEE P7000 series addresses issues at the intersection of technological and ethical domains. Similarly, the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems has produced the "Ethically Aligned Design" guidance which "sets forth scientific analysis and resources, high-level principles, and actionable recommendations. It offers specific guidance for standards, certification, regulation, and legislation for design, manufacture, and use of [AI] that provably aligns with and improves holistic societal well-being" [6]. We give an overview of the relevant P7000 series standards below, but note that only IEEE 7010-2020 is available, and the remainder are only Approved PAR.

Reference	Title
IEEE P7000	Model Process for Addressing Ethical Concerns During System Design
IEEE P7001	Transparency of Autonomous Systems
IEEE P7002	Data Privacy Process
IEEE P7003	Algorithmic Bias Considerations
IEEE P7004	Standard on Child and Student Data Governance
IEEE P7005	Standard on Employer Data Governance
IEEE P7006	Standard on Personal Data AI Agent Working Group
IEEE P7007	Ontological Standard for Ethically driven Robotics and Automation Systems
IEEE P7008	Standard for Ethically Driven Nudging for Robotic, Intelligent and Autonomous Systems
IEEE P7009	Standard for Fail-Safe Design of Autonomous and Semi-Autonomous Systems
IEEE P7010	IEEE Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being
IEEE P7011	Standard for the Process of Identifying & Rating the Trust-worthiness of News Sources
IEEE P7012	Standard for Machine Readable Personal Privacy Terms

Reference	Title
IEEE P7014	Standard for Ethical considerations in Emulated Empathy in Autonomous and Intelligent Systems

Table 19: The IEEE P7000 series

A brief summary of each of these standards are provided in [40]. The release schedule for the above standards is not known or available. With regard to the published IEEE 7010-2020, the standard aims to "establish wellbeing metrics relating to human factors directly affected by intelligent and autonomous systems and establish a baseline for the types of objective and subjective data these systems should analyze and include (in their programming and functioning) to proactively increase human wellbeing".

In addition to the above, the IEEE is developing standards in parallel targeting the development of robotics and autonomous systems [41]. The IEEE under the IEEE Robotics and Automation Society has formed a study group investigating "the feasibility of creating standards and performance metrics to measure robot agility, with the goal of enabling robots to be more productive, more autonomous, and to require less human interaction" [41]. Given the potential benefits of utilising autonomous robotics systems within operations affecting nuclear material, we outline standards under development (all which are only Approved PAR) for future reference below.

Reference	Title
IEEE P2817	Guide for Verification of Autonomous Systems
IEEE P2751	3D Map Data Representation for Robotics and Automation
IEEE P1872.1	Robot Task Representation
IEEE P1872.2	Standard for Autonomous Robotics (AUR) Ontology

Table 20: Relevant standards by the IEEE Robotics and Automation Society

These works seek to standardise the representation of, reasoning about, and verification of task knowledge in the robotics and automation domain.

A.1.3 ANSI/UL 4600

The ANSI/UL 4600 Standard for Safety for the Evaluation of Autonomous Products addresses fully autonomous systems that move such as self-driving cars, and other vehicles including lightweight unmanned aerial vehicles (UAVs). Although tailored towards autonomous transport, the standard "uses a claim-based approach which prescribes topics that must be addressed in creating a safety case. It is intended to address changes required from traditional safety practices to accommodate autonomy, such as lack of human operator to take fault mitigation actions" [6]. These topics are generally covered in a technologically neutral manner and may provide general guidance and insights applicable to autonomous operations affecting nuclear material. The relevant scope of UL 4600 includes

- safety case construction
- risk analysis
- design process
- verification and validation

- tool qualification
- data integrity
- human-machine interaction
- metrics and conformance assessment

Performance criteria and security-informed safety are not within the scope of the standard, although security is briefly addressed as a requirement.

A.1.4 Ethical and trustworthiness guidelines and other

There are further activities on the ethics of AI and its applications beyond international standardisation bodies. Notably, the High-Level Expert Group on AI in European Commission released the "Ethics Guidelines for Trustworthy Artificial Intelligence" report [12] which aims to provide a framework that promotes trustworthiness of AI/ML systems through a well-defined set of principles. These principles are now widely accepted and have been adopted worldwide by various industries, governments, and standards bodies. The framework achieves "trustworthy AI based on fundamental rights as enshrined in the Charter of Fundamental Rights of the European Union (EU Charter), and in relevant international human rights law". The principles are

- Human agency and oversight
  - including fundamental rights, human agency and human oversight
- Technical robustness and safety
  - including resilience to attack and security, fall back plan and general safety, accuracy, reliability and reproducibility
- Privacy and data governance
  - including respect for privacy, quality and integrity of data, and access to data
- Transparency
  - including traceability, explainability and communication
- Diversity, non-discrimination and fairness
  - including the avoidance of unfair bias, accessibility and universal design, and stakeholder participation
- Societal and environmental wellbeing
  - including sustainability and environmental friendliness, social impact, society and democracy
- Accountability
  - including auditability, minimisation and reporting of negative impact, trade-offs and redress

Further work on trustworthiness has been carried by 59 international co-authors from 29 organisations, including Adelard, to co-write "Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims" [14], which suggests 10 mechanisms for how AI developers can make more verifiable claims in three areas: institutional, software and hardware. The report emphasises the need to move beyond principles to a focus on mechanisms for demonstrating responsible behaviour.

The mechanisms provided in the report may provide direction for future research in addressing gaps identified in applying the existing UK nuclear regulatory guidance, including the ONR SAPs, TAGs, and Security Assessment Principles (SyAPs) to the application of AI/ML systems.

Finally, the Safety of Autonomous Systems Working Group (SASWG), which is convened under the Safety-Critical Systems Club (SCSC), have released guidance for "Safety Assurance Objectives for Autonomous Systems" [13]. The goal of the SASWG is to produce clear guidance on how autonomous systems and autonomy technologies should be managed in a safety-related context, throughout the lifecycle, in a way that is tightly focused on challenges unique to autonomy.

A.1.5 Standardisation landscape summary

We have outlined the landscape of international standards and guidance currently being developed to assure the implementation and deployment of AI, ML, and autonomous systems that may be relevant to operations affecting nuclear material. A more comprehensive summary of all international activities in AI from a more general viewpoint is provided in "ICT Standards and Ongoing Work at International Level in The AI Field - A Landscape Analysis" [42]. Given the limited availability of the published standards and guidance, the remaining tasks will thus only consider those available, that being

- ISO/IEC TR 24028:2020 Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence
- IEEE 7010-2020 Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being
- ANSI/UL 4600 Standard for Safety for the Evaluation of Autonomous Products [6]
- High-Level Expert Group on AI in European Commission — Ethics Guidelines for Trustworthy Artificial Intelligence [12]
- SASWG's Safety Assurance Objectives for Autonomous Systems [13]
- Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims [14]

Additionally, given that ISO/IEC/TR 29119-11 is under formal approval, if it's published within the timescales of Task 2, we will include it within the scope of our analysis.

The potential role of these standards and guidelines within the ONR regulatory guidance will be considered. In addition, we will consider which nuclear safety applications these standards may be applicable to (if any). We now provide an overview of each of these candidate standards and their applicability to the nuclear domain.

A.2 Detailed review of standards and guidelines

In this section we provide an overview summary of candidate standards which are available and their applicability to the nuclear domain.

A.2.1 ISO/IEC TR 24028:2020

This recently published technical report (TR) [43] focuses on the aspects of trustworthiness of AI systems, in particular, the potential factors that can impact trust. It defines trustworthiness as the ability to meet stakeholders' expectations in a verifiable way and notes characteristics that include reliability, availability, resilience, security, privacy, safety, accountability, transparency, integrity, authenticity, quality, and usability. We note that a stakeholder is defined as any individual, group or organisation that can affect, be affected by or perceive itself to be affected by a decision or an activity.

The TR documents some existing approaches that can improve and support trustworthiness and also some possible approaches to mitigating vulnerabilities in AI systems that relate to trustworthiness issues. In the TR, trustworthiness is considered in a similar way to security, as it is not a functional requirement of the

system, but a mechanism for meeting stakeholders' expectations. It specifically notes several different layers of trust, such as physical, cyber, social, etc.

The TR specifically lists some of the existing approaches and frameworks for trustworthiness and references additional standards which could be an additional useful resource to consider in further work.

The TR is broadly high level and provides an overview of several areas of interest, often introducing the topics and providing references to more in-depth documents. In Table 21, we identify the relevant areas listed in the report.

Section of TR	Relevant areas
Recognition of high-level concerns	• Responsibility, accountability and governance • Safety
Vulnerabilities, threats and challenges	• AI specific security threats (data poisoning, adversarial attacks) • Hardware threats to confidentiality and integrity • AI specific privacy threats • Unpredictability • Challenges to the specification of AI systems • Challenges to the implementation of AI systems • Challenges to the use of AI systems
Mitigation measures	• Transparency • Explainability • Controllability • Reducing bias • Reliability, resilience and robustness • Functional safety • Testing and evaluation (software V&V, formal methods, empirical testing, field trials) • Use and applicability

Table 21: ISO/IEC TR 24028 relevant areas

The TR also contains an annex on the societal issues for addressing trustworthiness of AI. This annex mostly points to examples of existing work, such as an IEEE analysis of issues around ethically aligned design of autonomous and intelligent systems [44].

A.2.2 IEEE 7010-2020

IEEE 7010 is a newly published standard promoting recommended practice, and specific and contextual well-being metrics for human well-being and the impact of autonomous and intelligent systems on it. The standard is grounded in the principle that businesses, governments, and individuals should aim to promote human well-being in the development of autonomous systems. The standard stays away from turning well-being into a single dimensional metric, but summarises it as [45]

"Well-being refers to what is directly or ultimately good for a person or population and depends on what is indirectly good for a person or population as well. Direct indicators for well-being capture people's reflection of how satisfied they are with their lives, their perceptions of their well-being, etc. While indirect indicators capture many important contributors and circumstances that lead to well-being, a direct indicator of well-being helps to understand overall well-being."

IEEE 7010 is intended for autonomous and intelligent system designers, developers, engineers, programmers, etc., in order to help in the following particular areas of interest:

- establishing a concept of human well-being in relation to autonomous and intelligent systems
- possible means for assessing impacts of autonomous and intelligent systems on human well-being over the whole lifecycle
- guidance for autonomous and intelligent systems development (in the context of promoting human well-being)
- informing risk mitigation strategies
- identifying stakeholders, intended and unintended users, uses, and their impacts on human well-being

Section of standard	Relevant areas
Well-Being Impact Assessment (WIA)	• Internal, user, and stakeholder impact assessment • Data Collection Plan and Data Collection • Well-Being Data Analysis and Use of Well-Being Data
Well-being domains and indicators	• Domain of environment • Domain of health • Domain of government
Annexes	• Integration of IEEE 7010 into existing processes

Table 22: IEEE 7010 relevant areas

Though the standard focuses on well-being and not safe functionality of autonomous systems, safe operation of the autonomous systems would make up part of societal acceptance, particularly if linked to the well-being of those in society. The standard also contains several checklists around WIA that potentially could be mined for principles related to well-being. An example from WIA Activity 1 Task 2 (user engagement):

"Were blind spots, potential biases, negative impacts, and other unknowns considered, including how risks and negative impacts to human well-being can be mitigated?"

In summary, IEEE 7010 may be less related to the functional safety or production excellence of devices containing AI/ML, but can help us frame principles around the wider use and acceptance of these technologies into society, in particular, those related to the well-being of people. We refer the reader the European Commission's "Ethics Guidelines for Trustworthy Artificial Intelligence" report [12] (Section A.2.5) as a more comprehensive and detailed framework that promotes trustworthiness of AI/ML systems, and is well established within the industry. However, IEEE 7010 and [44] are complementary and supplementary material.

A.2.3 ANSI/UL 4600

The ANSI/UL 4600 Standard for Safety for the Evaluation of Autonomous Products, although tailored towards autonomous transport, "uses a claim-based approach which prescribes topics that must be addressed in creating a safety case, which ensures that the standard takes an outcome-based approach. It is intended to address changes required from traditional safety practices to accommodate autonomy, such as lack of human operator to take fault mitigation actions" [8]. These topics are generally covered in a technologically neutral manner and may provide general guidance and insights applicable to autonomous operations affecting nuclear material; some example topics would be on their general guidance on safety cases (5.1) and fault modelling (6.2).

As well as providing examples, the standard also lists common "pitfalls" that could potential be mined for defeaters to the argument. Clause 8.5.5, "post-deployment changes to machine learning behavior shall not compromise safety", provides an example pitfall: "Modifying machine learning behaviour via reinforcement learning is prone to invalidating the safety case". This pitfall would certainly be a defeater focused on the changes made to ML behaviour invalidating safety and would need to be addressed within the safety case.

The standard also provides some clauses with examples, such as Operational Design Domain (ODD); but since it is an autonomous vehicle standard, the information in the examples may not be relevant. However, ODD is still relevant to the nuclear sector and clause 8.2.2, "the ODD shall cover relevant environmental aspects in which the autonomous item will be operating", would need to be re-envisaged to this domain.

The standard is aimed towards autonomous vehicles, thus it focuses heavily on neural networks which are commonly used in this domain (e.g., vision algorithms). One example is 8.5.2, "the machine learning architecture, training, and V&V approach shall provide acceptable machine learning performance", as although the clause itself is generic to all ML algorithms, it notes the number of layers in the network and hyper-parameters, which generally define a neural network model.

Table 23 presents the relevant scope of UL4600.

Section of standard	Relevant areas
Safety case construction	• Requirement of a safety case, including format • Sufficiency of goals, argumentation and evidence • Safety culture
Risk analysis	• Residual risks, including requiring a method for determining acceptable risk • Hazards • Risk evaluation techniques • Risk mitigation

Section of standard	Relevant areas
Design process	• Design and development process rigour • Product and development quality • Defect data • Dependability, including fault handling (detection and mitigation) and degraded operation • Redundancy • Robustness • Performance, such as time response
Lifecycle concerns	• Requirements and design validation • Supply chain • Field modification and updates • Operation
Verification and validation	• Verification and validation test approaches • V&V methods • V&V coverage • Fault model • Revalidation and change analysis
Tool qualification	• Tool qualification and COTS components • Tool risk mitigation • COTS and legacy risk mitigation
Human-machine interaction	• Risks associated with human interaction, including mitigation • Human interaction within the lifecycle (maintenance, commissioning) • Human contribution to operational safety, including responsibility • Communication (alarms, alerts etc.) • Training and awareness
ML and AI techniques	• Acceptable capabilities • Architecture • Acceptable data (including robust data validation) • Post-deployment changes to ML behaviour • AI techniques beyond ML

Section of standard	Relevant areas
Metrics and conformance assessment	• Run-time monitoring • Safety performance indicators • Metric definition • Metric analysis and response • Conformance assessment • Conformance monitoring
Cybersecurity	• Security planning and processes • Data integrity

Table 23: UL4600 relevant areas covered

Performance criteria and security-informed safety are not within the scope of the standard, although security is briefly addressed as a requirement. We are not proposing to review all sections in depth, as some are related more to general systems engineering rather than to the specific use of AI/ML technologies. However, we would consider any relevant impact of AI/ML on these areas. The standard has a table of clauses, which would be useful for us to quickly identify the most relevant clauses.

A.2.4 SASWG – Safety Assurance Objectives for Autonomous Systems

This publication [13] provides guidance for the safety assurance of autonomous systems under both SASWG and SCSC. It is aimed at multiple stakeholders from developers of autonomous systems to safety engineers and regulatory authorities.

The guidance focuses on three different levels of abstraction for autonomous systems:

- compute-level – implementation of AI/ML at the software and computational hardware level
- autonomy architecture level – integration of computations can be integrated into a system or platform
- platform-level – the final autonomous entity, its goals, and environment (heavy focus on requirements)

We believe that the guidance in the compute and autonomy architecture levels would be most in scope for our review. The compute-level, focusing on creating and verifying autonomous algorithms, could be useful for PE assessments of devices containing such components, and the architecture level for integrating AI/ML compute components into a wider device/system or platform, addressing implications for assuring system level attributes.

Some exclusions of note in the guidance:

- no consideration of criticality levels of functional safety (such as Safety Integrity Levels or Development Assurance Levels) in this version of the publication
- the role of domain specific certification and the challenges of this area are excluded
- the publication is heavily focused on machine learning types of AI, in particular neural networks, so gaps may be present for other types of algorithm

A.2.4.1 Compute-level

The table below lists the compute-level framework objectives of note.

Guidance principle	Relevance
COM1-1: Data is acquired and controlled appropriately	Data sets are very important in ML approaches to support the argument that the ML-produced algorithm has the correct behaviour. This extends to training, testing and verification data sets. Quality of ML-produced algorithms would probably need to be investigated in PE assessment.
COM1-2: Pre-processing methods do not introduce errors	Raw input data usually requires pre-processing to make it viable for ML algorithm inputs; this has the opportunity to introduce errors if mishandled. However, it is also an opportunity to detect missing or invalid input data.
COM1-3: Data captures the required algorithm behaviour	Ensuring the training data set captures the requirements of the specific algorithm's behaviour is an important part of development. The learning encoded from a training data set does not translate or map explicitly to the behaviour requirements, so an argument is needed as to why a particular set of data is appropriate for a set of specific algorithmic behaviour.
COM1-4: Adverse effects arising from distribution shift are protected against	The training data set does not differ from the operational input by a statistically meaningful way. The operational domain is captured in the training information so that the algorithm can operate safety.
COM2-1: Functional requirements imposed on the algorithm are defined and satisfied	Evidence will be required to show that the ML algorithm satisfies the functional requirements attributed to it. Issues of traceability from verification and testing to functional requirements could be difficult to overcome.
COM2-2: Non-functional requirements imposed on the algorithm are defined and satisfied	Performance, robustness and other requirements will also need to be satisfied. If the ML algorithm and model is complex, time response or computing requirements may be difficult to satisfy.
COM2-3: Algorithm performance is measured objectively	Measuring the performance of individual specific algorithms in a meaningful objective way can be difficult for AI/ML systems, with many performance measures in industry but minimal consensus. Specifying the measures and acceptance criteria for testing candidate algorithms is also challenging.
COM2-4: Performance boundaries are established and complied with	Constraints on what the algorithm can expect (inputs/outputs) and handle (valid values) are known and have been considered in development; else the algorithm may give unphysical/unsafe answers or its behaviour may be unpredictable.
COM2-5: The algorithm is verified with an appropriate level of coverage	An appropriate level of test coverage has been performed and the test case coverage has been justified. Often edge cases are hard to identify and specify in testing; so some argument of completeness should be made, but it may be infeasible in some cases to test every possible scenario.

Guidance principle	Relevance
COM2-6: The test environment is appropriate	Confidence in the test environment is required to ensure adequate assurance in the test results themselves. Additionally, validity arguments that are representative of the real world should be made particularly if it is a simulated environment.
COM2-7: Each algorithm variant is tested appropriately	If an algorithm has different variants based on different functions then all variants should be tested and correctly handled. Any testing carried over from one variant to another should be justified.
COM3-1: An appropriate algorithm type is used	There are many different types of algorithms. A justification for the selected one should be presented, and empirical arguments are likely to be required.
COM3-2: Typical errors are identified and protected against	As in traditional software, typical defects can be avoided through development procedures and processes that provide a base level of confidence in the development practices. Since AI/ML is an emerging field, they should be regularly reviewed as industry practices mature.
COM3-3: The algorithm's behaviour is explainable	A lack of traceable contributions throughout some black box algorithms leads to a lack of understanding on how a specific piece of algorithm behaves and its contributions to the final output. Furthermore, it's currently not feasible to extract a meaningful explanation of why the algorithm took a specific decision. This weakens predictability of how the algorithm is expected to behave in untested scenarios and raises questions about transparency and accountability.
COM3-4: Post-incident analysis is supported	Being able to investigate and learn from past incidents is vital to improving the safety of relatively new and immature autonomous systems. There should be information available to investigate failures to support post-incident analysis such as I/O, data, state information and any relevant environment data.
COM4-1: The software is developed and maintained using appropriate standards	A high level of quality in the software and any supporting libraries/toolkits should be maintained, else faults in this software could undermine any assurance argument. Traditional practices and existing safety-critical software development standards can help to address these problems and prevent typical errors.
COM4-2: Software misbehaviour does not result in incorrect outputs from the algorithm	Protection in the software should be built in to detect failures and misbehaviour. If these happen inside the AI/ML algorithm, it is not guaranteed that they would be detectable, so steps should be taken to limit this possibility.
COM5-1: Appropriate computational hardware standards are employed	Modern AI/ML software sometimes is run on more specialist hardware (GPU, TPU) to improve the performance (often carrying out many parallel computations at once). If novel technology is used it should be assured to the same level as more traditional digital systems hardware.

Guidance principle	Relevance
COM5-2: Hardware misbehaviour does not result in incorrect outputs from the algorithm	As with COM4-2, failures and misbehaviours of the hardware should be detectable. Furthermore, training is not typically performed on the operational platform due to its computing-intensive nature; therefore, assurances should ensure that any model optimisations for hardware (quantisation, pruning, clustering) does not affect its performance or behaviour.

Table 24: Compute-level principles

A.2.4.2 Architecture level

The principles are less directly related to autonomous components and more aimed at fulfilling system level properties, so we have reviewed the system property projections documented in the guidance.

System property	Relevance
Tolerance	This property focuses on the tolerance to faults and failures related to the autonomous component at the autonomous architecture level. It includes faults related to • invalid input handling • monitoring various aspects of the AI/ML component and health of sub-systems • confidence in the outputs Some aspects can be covered using traditional engineering techniques such as defence in depth or diversity. Measuring confidence in ML outputs is an ongoing research field that would require the creation and implementation of new techniques.
Information provision	This provision ensures that the autonomous architecture records and maintains the required information for relevant stakeholders. • information related to the operating environment – e.g. to support proactive maintenance activities • facilitating post-incident analysis • logging to facilitate further AI/ML algorithm development The provision only ensures that the information is available, not how it is used. This provision should be met to enable any post-incident analysis and to help prevent system failures.
Adaptation	This projection focuses on the management of changes to the algorithm after the initial operational use. Some algorithms utilise learning in the field and continuous learning to improve the algorithm. We would expect that any changes made to the algorithm should be verified and validated to ensure that the new version is safe. Therefore, learning in this manner may not be suitable for these types of systems. The change management processes should cover how versions of the algorithm sub-component are updated once in operation.

Table 25: Architecture-level projections

A.2.5 European Commission — Ethics Guidelines for Trustworthy Artificial Intelligence

The protection of human lives is paramount to the operation of systems governing nuclear material, as a system's failure or malfunction may result in: death or serious injury to people, loss or severe damage to equipment and property, and environmental harm. Considering the High-Level Expert Group on AI – Ethics Guidelines principles would provide a robust foundation to ensure that safety is prioritised against the potential uncertain outcomes of autonomous systems deployed in nuclear systems.

Although the principles are important notions to consider, they do not address technological feasibility, development, safety, or security of AI/ML systems. The noted technical robustness and safety attributes are only a few of the properties that must be considered to build trustworthiness in the behaviour of an AI/ML-based system. Furthermore, the attributes themselves are not well defined to allow their use in safety justification. Below, we adopt some of these principles, to form a more coherent set of principles that not only address socio-technical criteria for AI/ML-based safety systems, but also security and dependability attributes which would promote the assurance of such systems to be potentially deployed within the nuclear domain.

A.2.5.1 Need and intention

The aim of this principle is to ensure that the use of AI and ML are justified, and that alternative feasible technological solutions have been considered. That is, the benefits of the use of AI/ML outweigh the risks of the failure modes of AI relative to other solutions. Non-AI/ML solutions should be used whenever possible to mitigate the risks of their accuracy and uncertainty.

The intention of the use of AI/ML system must also be considered, and the use of an AI/ML system must not be malicious and should be rooted in fundamental rights, such as those considered by EU Treaties, the EU Charter, and international human rights law. These include respect for human autonomy, prevention of harm, fairness, and explicability. These attributes may be at odds with each other, thus further analyses may be required to justify the prioritisation of each principle.

A.2.5.2 Available skills

Development and maintenance of AI/ML systems require highly specialised expertise. In general, it is difficult to recruit ML specialists given their demand. A lack of expertise could lead to a poorly developed AI/ML system that may compromise all principles. It thus must be the case that developers of these systems demonstrate sufficient training and expertise in the subject matter. Special care must also be taken in how third party software is adapted, and the risks associated with their use must be well understood.

A.2.5.3 Accountability (and transparency)

Fields such as Explainable AI (XAI) have been attempting to build and understand AI/ML systems whose decisions are amenable to be understood and deciphered by humans. However, this is still an open research challenge that may not be overcome in the near future. Alternative mechanisms must thus be put in place to ensure responsibility and accountability for AI systems throughout the development lifecycle, including pre- and post-deployment. This includes developing a system which is auditable, and has fail-safe mechanisms that allow for immediate response and interference to any negative consequences which may arise out of the behaviour of the AI/ML system. Personnel responsible for the system's traceability and response should be identified.

A.2.5.4 Data fairness, bias, and adequacy

It has been consistently demonstrated that data sets utilised to train AI/ML systems are riddled with historic and implicit biases, due to either incompleteness or lack of inclusivity in training data sets. Such biases in training data often leads to indirect prejudice and discrimination against certain groups or people by the developed AI/ML system, exacerbating prejudice and marginalisation further. When selecting or creating training datasets, it must be the case that diversity, bias, ethics, privacy, and fairness are all considered.

A.2.5.5 Security-informed safety

Security-informed safety should be addressed at all stages of the development lifecycle, from conceptualisation, experimentation, and prototyping through to production. A security-informed hazard analysis should be undertaken during development. The hazard analysis should be reviewed periodically during operation or when a safety-related component has been updated or if additional threat and vulnerability information has been identified.

A.2.5.6 Diversity and defence in depth

Diversity should be considered within the construction of a system's architecture to reduce the trust needed in a single ML component. Independence of failures should not be assumed and failure correlation should be considered based where possible on experimental data. An architectural approach which limits reliance on sub-components of the system that need to be highly trusted (e.g., ML models) should be taken.

A.2.5.7 Dependability requirements

Although ML-based systems bring about novel issues to the frontier, dependability properties must still be investigated to ensure the safety and behavioural correctness of the system. There are a variety of ways in which the desired properties of a system can be classified, but the decided catalogue of behavioural attributes currently used are functionality, performance, reliability, operability, robustness, availability and security. Although these attributes are traditional in their meaning, new techniques must be devised against ML-based systems, as existing methods are not applicable.

A.2.5.8 Adaptation and change (resilience)

Stakeholders need to have confidence, before an ML-based system is deployed, in how it is going to adapt to changes post-deployment. This is particularly important when considering safety requirements. The future behaviours of ML-based systems should be assured systematically through Open Systems Dependability (OSD) deployed on the system's lifecycle.

A.2.5.9 Verification and validation (V&V)

There are numerous behavioural attributes that must be considered to ensure the integrity and correct behaviour of an AI/ML system, including functionality, accuracy, reliability, operability, robustness, and availability. However, given the nature of AI and ML algorithms, traditional V&V methods are not applicable, and special care must be taken in applying corresponding AI/ML techniques given their novelty and potential immaturity. It is thus crucial to strategise the use of V&V to identify the role of such methods and how they complement other approaches. This includes techniques such as performance metrics, formal verification and static analysis, and simulation.

A.2.5.10 Interaction and oversight

Human oversight should be possible to help ensure that an AI system does not undermine safety or human autonomy. That is, it should be possible to interfere with an AI/ML system through any stage of its operation to allow for manual oversight, if necessary. Various governance mechanisms and strategies exist that may be suitable for varying systems, such as a human-in-the-loop (HITL), human-on-the-loop (HOTL), or human-in-command (HIC). Additionally, those making use of the AI/ML systems should be given the knowledge or training regarding the interactions with the system, including comprehensive understanding of the failure modes requiring human interference or oversight.

A.2.6 Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims

This report suggests various steps that different stakeholders can take to improve the verifiability of claims made about AI systems and their associated development processes, with a focus on providing evidence about the safety, security, fairness, and privacy protection of AI systems. The paper focuses on ten mechanisms for this purpose – spanning institutions, software, and hardware – and makes recommendations aimed at implementing, exploring, or improving those mechanisms. These mechanisms are:

1. Third Party Auditing
2. Red Team Exercises
3. Bias and Safety Bounties
4. Sharing of AI Incidents
5. Audit Trails
6. Interpretability
7. Privacy-Preserving Machine Learning
8. Secure Hardware for Machine Learning
9. High-Precision Compute Measurement
10. Compute Support for Academia

We note some relevant conclusions regarding the gaps prevalent to the assurance of AI claims below:

- It is difficult for AI developers to address – and demonstrate that they are addressing – the "unknown-unknowns" associated with AI systems, including limitations and risks that might be exploited by malicious actors. Further, existing red teaming approaches are insufficient for addressing these concerns in the AI context. Organisations developing AI should run red teaming exercises to explore risks associated with systems they develop, and should share best practices and tools for doing so.
- AI systems lack traceable logs of steps taken in problem definition, design, development, and operation, leading to a lack of accountability for subsequent claims about those system's properties and impacts. Standards setting bodies should thus work with academia and industry to develop audit trail requirements for safety-critical applications of AI systems.
- It's difficult to verify claims about "black box" AI systems that make predictions without explanations or visibility into their inner workings. This problem is compounded by a lack of consensus on what interpretability means. Organisations developing AI and funding bodies should support research into the interpretability of AI systems, with a focus on supporting risk assessment and auditing.

The noted mechanisms could potentially be deployed and further explored within the existing UK nuclear regulatory guidance to address gaps where current techniques are not sufficient to assure AI and ML

systems. However, the safety claims which these techniques support must be examined to determine compatibility with the ONR SAPs and TAGs.

Appendix B
Metrics and AI/ML performance

In this Appendix, we give a brief overview of some common probabilistic metrics used to evaluate an AI/ML system's performance. These are summarised in Table 6 in Section 3.3. We then discuss other arguments that can aid in building confidence in AI/ML reliability, for example using model-based approaches and Conservative Bayesian Inference. The Appendix is partially based on [39].

B.1 Binary classifiers

Binary classifiers return a TRUE or FALSE output, for example identifying if a system is in a dangerous state. Generally, they work by producing some numerical value (usually between 0 and 1), which is converted into a binary outcome by a threshold value separating the TRUE and FALSE states. The threshold is chosen to optimise the performance of the model (which may depend on the model's application, and the relative risk of making false positive and false negative decisions).

The simplest metric one might use would be *accuracy*, defined as the proportion of the inputs on which the model gave the right answer. This is simple to calculate and relatively easy to understand. However, accuracy does not take account of the distribution of the data on which the test is performed. If the model is being trained to recognise a rare phenomenon, which occurs in only 1% of cases, we can achieve 99% accuracy by simply answering "no" to all inputs. More nuanced measures, which allow for different distributions of input data, are therefore much more commonly used.

B.1.1 Receiver operating characteristic (ROC) curves

ROC curves are a way to estimate how effective a binary classifier algorithm is.

If a threshold value was chosen that was very close to zero, almost all observations would be classed as TRUE. As a result there would be a very high rate of false positives, but a very high rate of true positives too. As the threshold value is raised, the number of false positives decrease, but so does the true positive rate. Plotting the false positive rate against the true positive rate parametrically as a function of the threshold value generates a Receiver Operating Characteristic (ROC) curve. The area under this curve (AUC) gives a measure of the ability of the ML method to correctly distinguish the TRUE and FALSE states. Random guessing achieves an AUC of 0.5, whilst a perfect classifier gives an AUC of 1. An example ROC curve is shown in Figure 7. Note that an AOC value corresponds to a measure across all possible threshold values – it is therefore not specific to a particular implementation of the threshold.

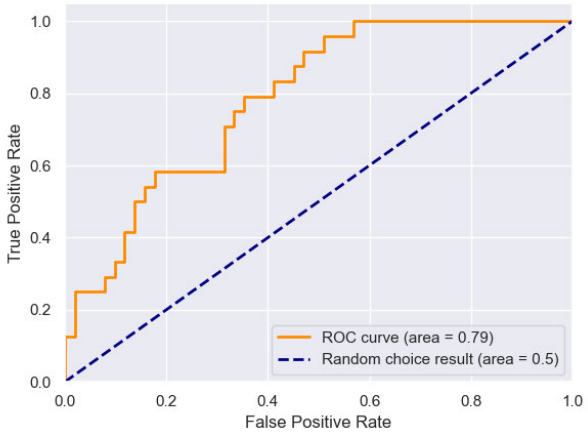

Figure 7: Example ROC curve

B.1.2 Precision and recall

Precision and recall measure the performance of a binary classifier for a particular threshold value. For a given set of data, there will be a number of true positives (TP), false positives (FP), false negatives (FN) and true negatives (TN), illustrated in Figure 8.

The precision, p, is defined as the proportion of the TRUE responses which were correct:

$$p = \frac{TP}{TP + FP}$$

This is illustrated in Figure 9 (left). Precision is a useful metric in situations where the consequences of a false positive are particularly severe. Examples would include facial recognition for access control, or an autonomous vehicle determining whether it is safe to turn right.

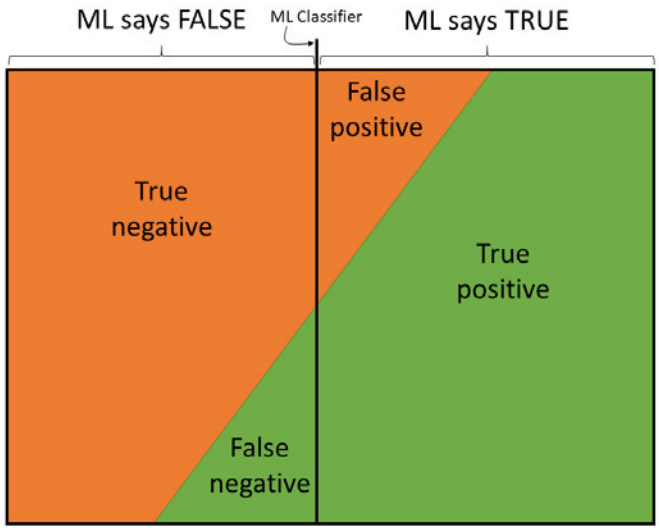

Figure 8: A binary classifier showing the ML classifier (vertical line) and the ground truth (red for FALSE, green for TRUE)

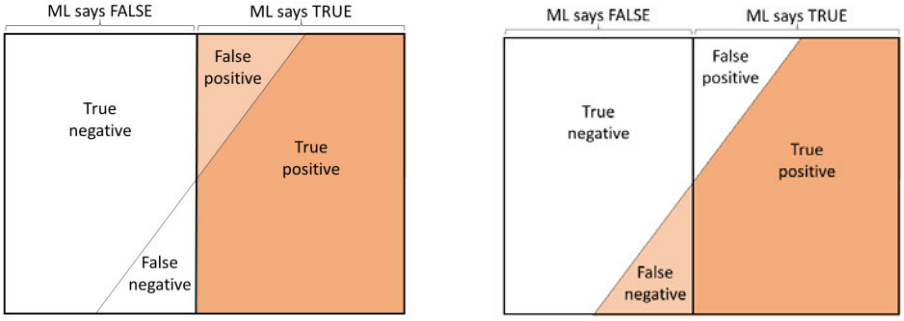

Figure 9: Precision (left) is the fraction of true positives, of all the detected positives and Recall (right) is the fraction of true positives, of all the ground truth positives.

Recall is a measure of "how good the ML is at identifying the property". Formally, the recall, r, is defined as the proportion of genuine "yes" instances identified by the ML:

$$r = \frac{TP}{TP + FN}$$

This is illustrated in Figure 9 (right). Recall is a useful metric where it is important to avoid false negatives. Examples include determining whether an MRI scan is potentially cancerous, or an autonomous system determining whether there is another vehicle in its path.

B.1.3 F_1 score

The F_1 score combines the precision and recall values through their harmonic mean:

$$F_1 = 2\,\frac{p \cdot r}{p + r}$$

A perfect classifier has an F_1 value of 1.

The F_1 score can be generalised to account for the relative importance of precision and recall in a particular application. In an application where recall is considered β times more important than precision, the performance can be measured using:

$$F_\beta = (1 + \beta^2)\,\frac{p \cdot r}{\beta^2 p + r}$$

Measuring the value of F_β would provide a measure of the performance of a classifier weighted suitably towards recall or precision.

B.2 Object detection

Performance measurement in object detection and classification is somewhat more complex than a binary classifier. Aside from the addition of more categories, which can be handled by the above measurements for binary classifiers, there are two further complications:

- the ML can (and is expected to) make multiple predictions on a single image
- predictions can have varying degrees of correctness, depending on how closely the location matches the object in question

In the following sections, we discuss this second point in detail.

B.2.1 Intersection over Union (IoU)

The "correctness" of a location prediction is usually measured in terms of intersection over union (IoU). The IoU of a prediction is defined as the area of the intersection of the predicted region and the true region, divided by the area of their union, and so is in the range 0 to 1. Using the IoU does omit some information about the accuracy, e.g. predictions which are twice the actual size, and half the actual size will be equally correct, but it is the generally accepted means of determining whether a prediction is correct. Some alternatives or modifications to IoU have been proposed [46].

The suitability of IoU as a means of determining whether a prediction is correct may depend on the intended application. Figure 10 provides examples of predictions (in green) compared to the ground truth (in black), all of which have IoU of 0.5. It can be seen that errors in size are penalised less than errors in position.[2] This is not necessarily intuitive and, in certain applications, may not be desirable. Indeed, it has been found to be difficult to train humans to determine the difference between IoU values between 0.3 and 0.7 [47].

[2] The first prediction is 41% too large in both dimensions, whereas the second prediction is displaced by 18% in each dimension.

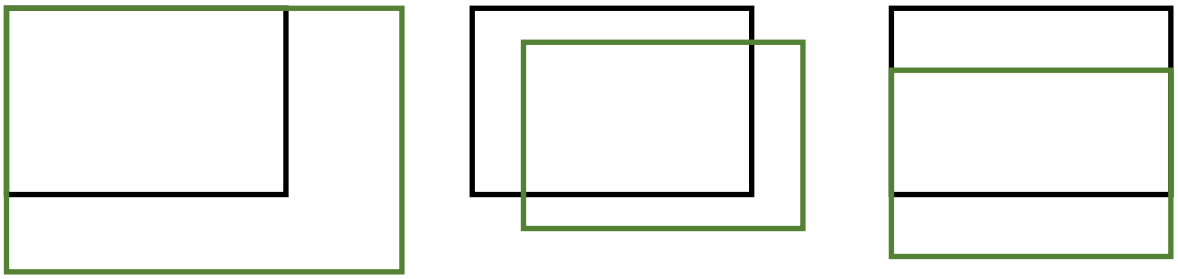

Figure 10: Example predictions with IoU = 0.5

An average precision metric can be formulated, based on IoU data, for example predictions are considered correct if the IoU of the prediction is above some fixed value *k*, and the average precision is taken by altering the threshold required to make a prediction. For an IoU value *k*, this is denoted AP^k. In general, higher values of *k* result in a lower average precision. The precision $AP^{0.5}$ is a common metric that has historically been used for measuring the performance of object detection algorithms. The overall average precision, also named "mean average precision" is the average of AP^k for values of *k* between 0.5 and 0.95 in intervals of 0.05. This average precision measure is also measured when only considering objects of a particular size – in general object detection algorithms perform better on larger objects.

Another measure used to measure performance is average recall. To measure average recall, the algorithm is permitted to make a fixed number, *d*, detections on each image. The recall is then measured for each category at a number of different IoU thresholds, and the average taken to produce the average recall AR^d. Given the relative consequences of a false negative and a false positive in many situations in autonomy, this metric may be more useful than average precision. However, the goal of both average precision and average recall in object detection is to allow for comparisons between algorithms, and they provided relatively limited information on the absolute performance for a particular configuration.

The best performing algorithms for the COCO dataset currently recorded on *cocodataset.org* achieved an $AP^{0.5}$ of 0.77 and an AR^{100} of 0.70.

B.2.2 Object tracking

Object tracking aims to identify and track the trajectories of multiple objects in a sequence of images (i.e. a video). Object detection is clearly a key element of object tracking, and most algorithms for object tracking consist of an object detection algorithm and a tracking algorithm to process the predictions made by the object detection. In contrast to object detection as considered above, objects generally do not need to be detected in every image to be tracked.

Common metrics measured to evaluate the performance of object trackers include

- Mostly tracked (MT): objects which were tracked for more than 80% of the time they were in frame.
- Mostly lost (ML): objects which were tracked for less than 20% of the time they were in frame.
- Identity switches (IDS): instances where the same object is identified as different objects in different frames.
- Fragmentations (FM): the number of gaps in trajectories of the same object.
- False positives (FP): the number of false positive detections.
- False negatives (FN): the number of false negatives.
- Multiple object tracking accuracy (MOTA): the total number of false negatives, false positives and identity switches, divided by the total number of objects across all frames.

- Multiple object tracking precision (MOTP): the average dissimilarity between each identified object and its true position.

For understanding the performance of object detection algorithms, the false positive and false negative metrics are likely to be the most useful. The MT and ML metrics provide information on what proportion of objects were consistently identified or consistently missed, respectively. An AI/ML product may not need to detect an object in every single frame, and so MT and ML may be suitable metrics for assessing what proportion of objects are detected frequently enough to be tracked.

The remaining metrics focus on specific aspects of tracking and evidence of the performance of object detection and tracking against these metrics is likely to be relevant to a safety case. The consequences of inconsistent identification and tracking of objects can be severe. FM and IDS provide metrics which could be applied to the overall system for object detection and tracking (which may consist of several different sensors, including cameras, LIDAR, etc.).

Multiple object tracking is an active area of research, and as a result there are a number of common datasets available to provide performance benchmarks. In general, the annotated objects to be detected are either pedestrians (for example, the MOT16 dataset) or vehicles (for example, the UA-DETRAC dataset). Datasets have been collected using both stationary and vehicle-mounted cameras.

To provide an overview of the current state of object tracking, Table 26 contains some data taken from a recent object tracking survey [48] on the performance of various trackers on the MOT16 dataset according to the metrics described above. The MOT16 test set contains a total of 759 tracked objects across 5,919 frames, with a total of 182,326 individual objects.

To separate the performance of trackers from object detection, the MOT16 dataset is also published both as simple videos, and as videos annotated with the objects detected by a fixed object detection algorithm. Following [48], we refer to these as private detection and public detection respectively. The difference in performance between the two serves to highlight the influence of object detection performance on object tracking performance.

Algorithm	Detection	MOTA	MOTP	MT	ML	FP	FN	FM	IDS
Customised Multi-Person Tracker	Public	49.3	79.0	17.8	39.9	5,333	86,795	535	391
Non-uniform hypergraph learning based tracker	Public	47.5	--	19.4	36.9	13,002	81,762	1,408	1,035
Person of Interest	Private	68.2	79.4	41.0	19.0	11,479	45,605	1,093	933

Algorithm	Detection	MOTA	MOTP	MT	ML	FP	FN	FM	IDS
Lifted Multicut and Person Re-identification (LMPR)	Private	71.0	80.2	46.9	21.9	7,880	44,564	587	434

Table 26: Performance on MOT16 benchmark dataset

The FP and FN metrics allow us to compute precision and recall scores to provide some comparison with metrics for object detection algorithms. In the case of LMPR, the corresponding precision is 0.946, and recall 0.756. This recall is similar to the recall achieved on the COCO dataset by the best performing detectors, but with substantially higher precision. The increased precision is probably due to the more limited set of object that are being detected, as is the case for traffic lights.

It is notable that the proportion of tracks which were mostly lost is 21.9%, which is not substantially lower than the complement of the recall, i.e. the proportion of objects which were not detected. We can make similar observations for the other trackers in Table 26. This raises the possibility that those objects which are missed by a detector in an image are consistently missed in subsequent images, which casts doubt on any claimed performance gain by detecting objects across multiple images.

B.3 Approaches to increasing reliability claims

B.3.1 Model-based approaches for object detection

Even in ideal circumstances, the best object detection algorithms rarely achieve precision and recall scores above 0.9 when predictions are made on a single image. Some improvements in this performance can be made by using an ensemble of similar algorithms, or by including additional information such as GPS data in the detection but this is still significantly below the accuracy that would typically be required of a safety-critical system.

In general, cameras and object detection algorithms, and other types of sensor, are used by an autonomous system to construct a model of its environment. In the case of object detection, this is a model of the physical environment, including different types of objects, their locations, motion, inferred intention, etc.

An improved approach to reasoning is to require high confidence in the output of the resulting model, rather than in any individual sensor. One such approach is the predictive processing discussed in Section 2.3.1. The predictions made by the sensors are compared with predicted sensor outputs based on the model of the environment. If all sensors make predictions with high confidence, and these predictions only differ slightly from those expected based on the model, then we have high confidence that the model is an accurate representation of the environment.

In the case that there is a large discrepancy between the prediction of the model and the prediction of a sensor, then this must be resolved. If the error is only present in one sensor, or one group of related sensors, e.g. all front-facing cameras, then it may be determined that this is a sensor error, and the model can be updated based on the data from other sensors. If such errors persist, then it may be due to some external cause, e.g. the cameras are being dazzled by the sun. Finally, if a large number of sensors provide

a prediction error, then it is likely that the world has not developed as predicted by the model, e.g. a new object has appeared, or an object has moved in an unexpected way. In this case, the model must be updated to reflect the new data.

Using such a predictive processing model presents a possible approach to arguing that an autonomous system's interpretation of its environment is accurate with a sufficiently high degree of confidence, without requiring a potentially infeasible level of performance from an individual sensor. If a sufficiently high level of accuracy, e.g. measured by precision and recall, can be achieved then these can support a higher level of reliability in the model. For example, if five independent sensors each have precision and recall of 0.9, then the "majority verdict" has precision and recall of 0.991.

Demonstrating the reliability of the model in this way is similar to increasing performance using ensembles. In particular, we require evidence that the failure of sensors is independent. There could be many causes of systemic failures, such as incomplete training data, or sensors having reduced performance in the rain. The impact of such systemic failures on the performance of the model will depend on the precise nature of the failure. Reduced performance in rain will affect the accuracy of the sensor data, however the errors are unlikely to be consistent, and these errors are likely to be identified. On the other hand, if an object is not seen in the training data, then the sensors may consistently fail to identify it, leading to an incorrect model. To support a claim for independence of the sensors, we would require evidence identifying any potential sources of systemic errors, and arguments as to why they do not occur.

The predictive processing model also allows for additional reasoning about the model itself, which can increase our confidence further. One such method is to keep track of any uncertainties in the model, which can then be resolved by further observations. For example, if an object is identified but the sensors cannot determine whether the object is a bicycle or a pedestrian, this uncertainty can be maintained within the model (or as multiple models) until the object can be identified from further observations. Assurance of such multiple models would likely take the form of showing that the true state of the world is represented within the space of possible models with sufficient confidence, and that the actions planned and taken by an autonomous vehicle are safe in all possible models.

The model would also be able to track objects which are partially or totally obscured. This includes both previously detected objects which have since become hidden, and potential objects in areas which are not visible to the sensors, e.g. if a parked vehicle is blocking the view of the road. Doing so would be necessary to ensure the accuracy of the model. Some work on detecting and representing the properties of objects and their relative positions in an image, and answering queries regarding these, has been performed in [50]. Using a combination of YOLO and answer-set programming (ASP), this algorithm was able to answer 93.7% of queries correctly. The vast majority of the errors were caused by incorrect object detection with YOLO, and we assume that the small number of errors arising from incorrect parsing of the natural language queried could be eliminated. It is not clear how well this performance will translate to the objects of interest to AVs, but it is plausible that ASP could be used to develop systems that can reliably reason about the relative positions of objects in a model.

B.3.2 Conservative Bayesian Inference

The high reliability claims required to have sufficient confidence in AI/ML systems often lead to prohibitively extensive or impractical testing requirements. New argumentation structures can help build confidence in reliability claims when operational data is limited; this might include testing in complex environments such as road testing an autonomous vehicle.

Conservative Bayesian Inference (CBI) [37] [38] provides a route to build confidence in a product using both operational data and lifecycle information, in a conservative way. This approach is consistent with the ONR

SAPs, under paragraph 191, where in the case of unavailable reliability data, the demonstration can include *"a review of precedents set under comparable circumstances in the past"*. CBI presents a quantitative method to incorporate reliability data with these precedents. Moreover, ERL.4 states: *"Where safety-related systems and/or other means are claimed to reduce the frequency of a fault sequence, the safety case should include a margin of conservatism to allow for uncertainties. "*, as such, CBI is appropriate due to its absolute conservatism. An example CAE structure for a CBI argument is shown in Figure 11.

CBI also formalises and quantifies the often used approach of estimating a reliability claim, then assuming a weaker reliability claim with higher confidence in subsequent analysis.

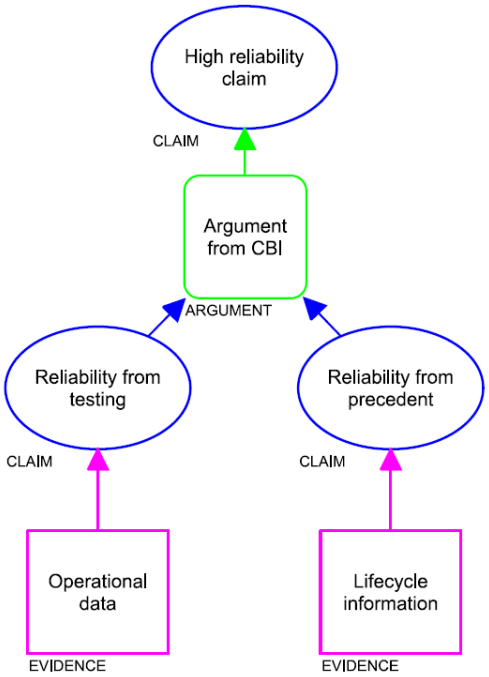

Figure 11: Example CAE structure for a CBI-based reliability claim

B.3.2.1 How Conservative Bayesian Inference works

Bayesian inference allows for updating a prior distribution with new information.

Let γ represent the system's unknown reliability. We wish to determine an accurate distribution of γ based on a prior estimate of this distribution, and some new reliability data.

First we must determine a prior distribution, $P(\gamma)$, which can take the form of any valid probability distribution. This is our estimate of the distribution of γ based on lifecycle information.

We then observe a data point (or set of data points). These can take the form of operational testing data – this could be a failure rate of an autonomous vehicle on a test drive or a precision score for an algorithm tested on a data set, for example. Our prior distribution can be updated using Bayes theorem to give a posterior distribution. This posterior distribution is our new best guess for the distribution of γ, from which we are able to directly make claims such as "$\gamma < 10^{-6}$ with confidence 90%." This process is illustrated

graphically in Figure 12 – in this example the product can be more confidently categorised as SIL 3 in the posterior distribution.

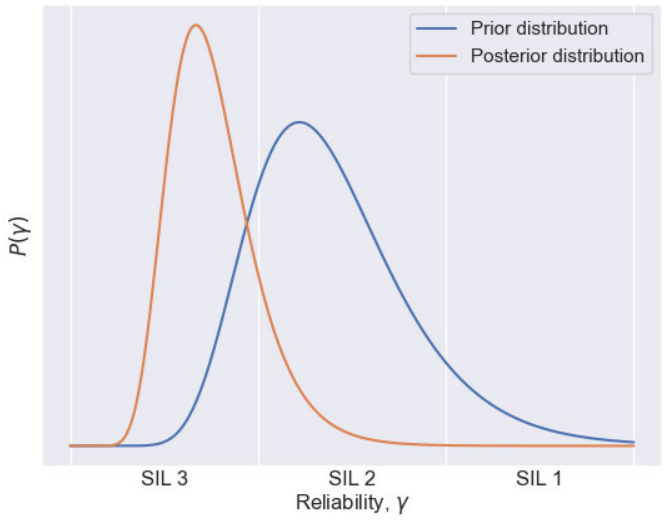

Figure 12: The use of Bayesian inference can increase confidence in a product

This approach may not be practical if a good description of a prior distribution is not available. In the case of only limited prior knowledge, Conservative Bayesian Inference can be used.

To perform a CBI argument, we only need *partial* knowledge of the prior distribution. This could be expressed as a maximum of the prior mean, or a confidence bound on γ, or some combination of these. For example, we may have confidence θ, that γ is less than ε. There may also an absolute minimum value of γ, δ (In the example of an autonomous vehicle, this might represent the rate of catastrophic hardware failure whilst driving). As such, our prior is constrained by

$$P(\gamma < \varepsilon) = \theta \;\; AND \;\; P(\gamma > \delta) = 1 \;\; AND \;\; P(\gamma < 1) = 1$$

There are an infinite number of possible priors that satisfy these two conditions. Two examples are shown in Figure 13.

Each of these priors could be used in a Bayesian Inference analysis to determine the posterior distribution of γ given the observed data. In CBI, we choose the prior that, when converted into a posterior, gives the bleakest possible prediction for whatever property we are interested in. This is guaranteed to give the most conservative value of a property given the prior constraints. Note that the most conservative prior is dependent on exactly what property of the posterior we want to understand.

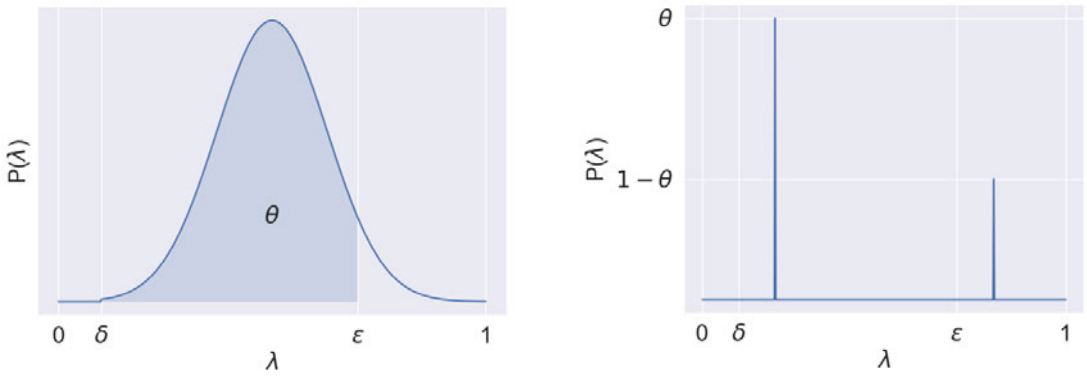

Figure 13: Two example prior distributions that satisfy the constraints described in Section B.3.2.1

Appendix C
Monitors and defence in depth

Given that ML-based components, especially perception systems, are difficult to assure, approaches are needed to reduce the assurance burden and allow their use. Consider engineered complex system architectures, which are used to limit parts of the system that need to be highly trusted: safety and security protection is provided in a simpler system or safety monitor that detects when a system is close to being in an unsafe or insecure condition, and acts accordingly. In this section, we thus address the impact of the use of safety monitors within an assurance case. A safety monitor architecture is common across different disciplines (e.g., aircraft, railway systems, nuclear power plants, etc.) and is proposed as a standardised approach in the air domain [51], and more generally, for cyber physical systems [52].

We first consider near-term monitor style architectures, followed by a discussion on their disadvantages and why they may not be sufficient. We then introduce the notion of more cognitive predictive and processing-based approaches, and additionally consider the alternative approach of neuromorphic computing. The appendix draws heavily on our previous work [39].

C.1 Monitor-based architectures

The architecture of any autonomous systems should continue to use familiar methods for achieving high reliability and trustworthiness: the use of redundant hardware, the use of guard or monitor-based architectures, and the provision of defence in depth. For the latter there may be system level considerations such as human-factor interactions, and employing diversity at sensor level (use of LIDAR, cameras, radars, etc.).

Safety monitors can vary in sophistication from comparison between diverse sensors (e.g., comparison of LIDAR measured distance with that from a stereo camera) to a monitor implementing a complex set of equations and constraints (e.g., see Responsibility-Sensitive Safety (RSS) [53]). This architectural approach often seeks to reduce the trust needed in ML components by monitoring both the state of the environment and the vehicle. They can also monitor when an autonomous system is under stress, or in an error prone situation. It is not unlike the intrusion detection problem in security, where one tries to infer potentially dangerous behaviour from the complex system state and knowledge of threats. The Defence Advanced Research Projects Agency (DARPA) Assured Autonomy programme, for example, extends the safety monitor concept to include a dynamic assurance case, as monitors can be seen as a form of run-time certification that shifts the certification or assurance challenge from the design and development part of the lifecycle to operation [54].

We are particularly interested in how safety monitors can be used to gain the performance and, in some cases, safety benefits of deploying complex ML components, whilst mitigating the risks of using such technologies. One safety-monitor approach aims to support the assurance of an architecture that limits reliance on sub-components of the system that need to be highly trusted (e.g., ML algorithms). In Figure 14 we have adapted the safety monitor architecture of [51] to include both a safety monitor and a complex function monitor, implemented for an AI/ML-based system (note that we use the term AI/ML rather than the term Learning Enabled Component (LEC) used in [51]).

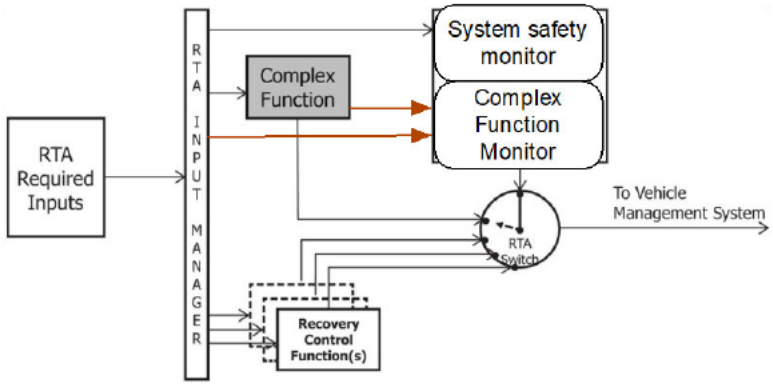

Figure 14: Safety monitor architecture

The recovery functions also must address a number of design challenges. One of the design challenges for an AI/ML monitor is illustrated in Figure 15:

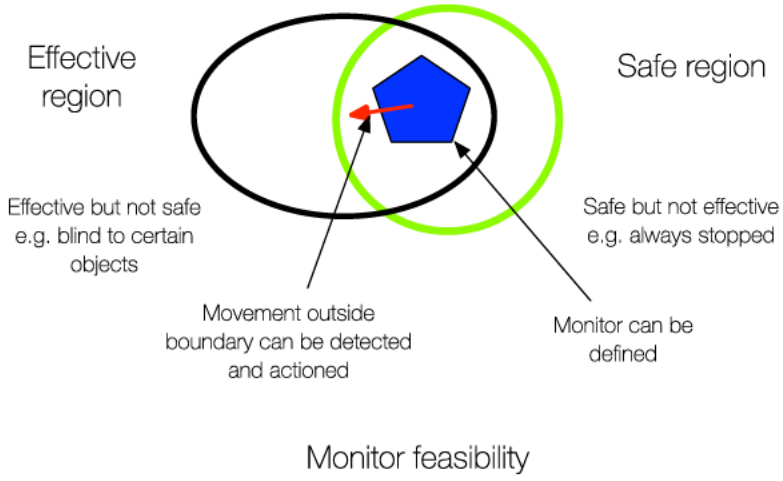

Figure 15: Monitor feasibility

In the design of this architectural strategy, a number of questions must be addressed:

- Is a monitor feasible – can a region be identified that is both safe and effective and can be described in terms that can be assured more readily and to greater extent than the AI/ML itself?
- Can transgression of the monitor region be detected and actioned by a recovery control function?

A number of approaches can then be taken, which we characterise as:

- *Environment Monitors:* monitor an ML system's input space where there are known performance issues – e.g., bad weather, flying upside down, etc.
- *Health Monitors:* monitor the ML system's internal state and identify states that might be "stressed" or indicative of a problem (e.g., monitoring activation patterns, resource utilisation, and simple tasks for which diverse measurement is possible).
- *Behaviour Monitors:* monitor the ML system's outputs and inputs to see if they violate bounds on specified behaviours or invariants.

Finally, the recovery strategies are very application specific. They may involve the use of other sensors in the case that the ML system has degraded performance, but will still allow for safe behaviour (e.g., moving to a minimum risk position or reduced performance while recovery is planned).

In practice, the architectures could be far more complex than shown in Figure 14. There is also the inverted architecture where AI/ML is used to learn the difficult tasks (e.g., how to conduct difficult and dangerous manoeuvres in an aircraft) and has the authority to take over from the more conventional operation. The monitors could thus become nested with an AI/ML monitor of a conventional monitor of an AI/ML-based sensor.