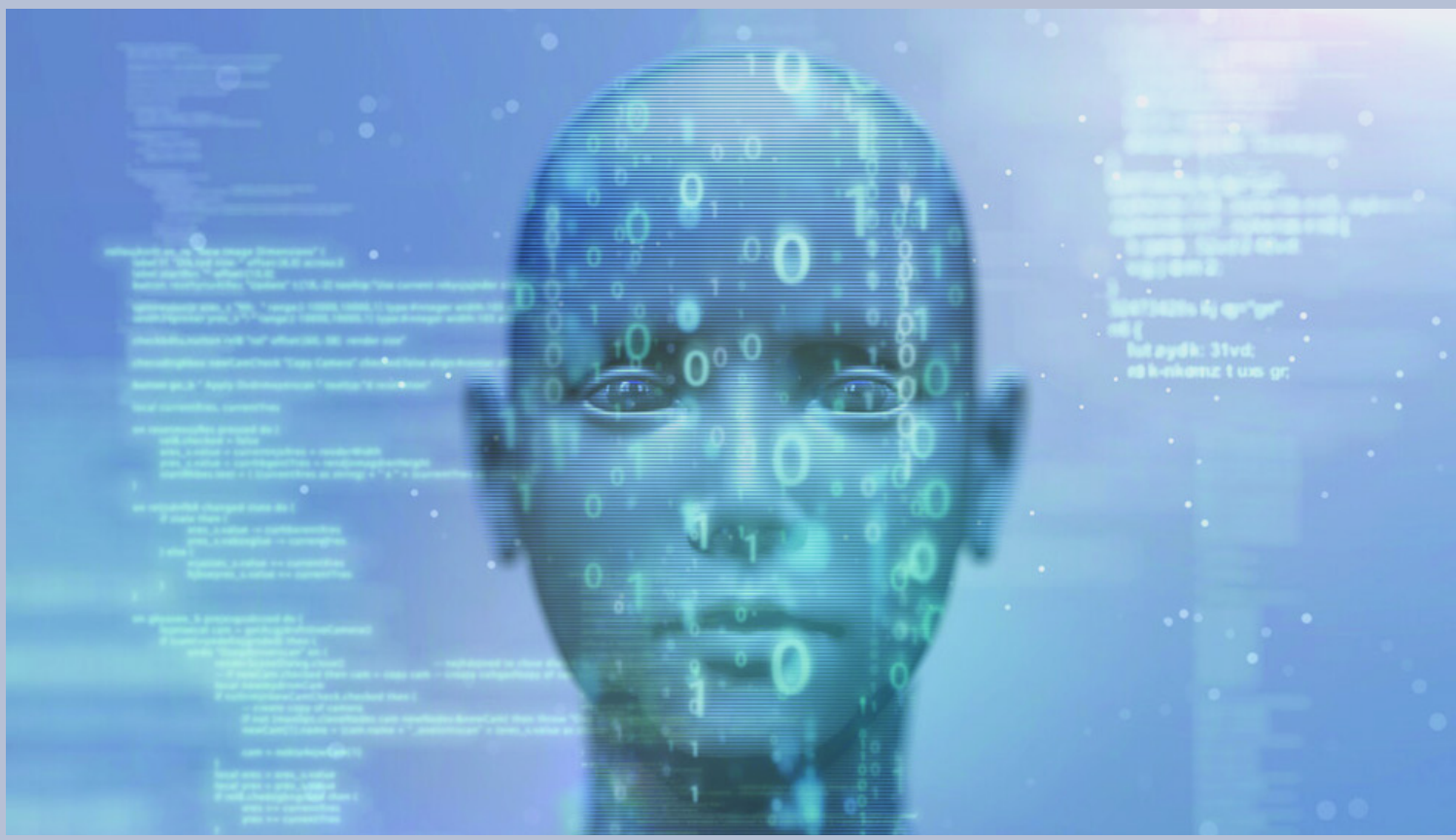
The Institution of Engineering and Technology

[dstl] The Science Inside

# The Application of Artificial Intelligence in Functional Safety



# Publication information

Published by: The Institution of Engineering and Technology, London, United Kingdom
The Institution of Engineering and Technology is registered as a Charity in England & Wales (no 211014) and Scotland (no SC038698).


First published 2024



The Institution of Engineering and Technology
Futures Place, Kings Way
Stevenage
Herts
SG1 2UA
United Kingdom

Copies of this publication may be obtained from:
PO Box 96
Stevenage
SG1 2SD, UK
Tel: +44 (0)1438 767328
Email: sales@theiet.org
https://electrical.theiet.org



Typeset in the UK by The Institution of Engineering and Technology, Stevenage

# Contents

# Contents

# Acknowledgements

In putting this publication together, the IET's Engineering Safety Policy Panel formed a working group from across various sectors and would like to acknowledge assistance from the following individuals:

**Technical committee and other contributors**

| | |
|---|---|
| Isaac Akintaro | Arup |
| Lourdes Aristondo | Multimatic Inc |
| Alec Banks (Dr) | Dstl |
| Russell Bee | IBM |
| Amélie Beedham | Thales UK |
| Audrey Canning | Virkonnen |
| Steph Carroll (Dr) | Frazer-Nash Consultancy Ltd |
| Diogo Costa | Minho University |
| Melanie D'Mellow | Animal Dynamics |
| Julia Downes (Dr) | Dstl |
| Steve Frost | Formerly Office for Nuclear Regulation Inspector |
| Andrian Harsono (Dr) | The Pirbright Institute |
| Lesley Martin (Dr) | Formerly Siemens Mobility |
| Rhod Morgan | Formerly HSE Inspector |
| Phillip Mulvana | Dyson Technology |
| Shrey Patel | Angiras Rasayan LLP |
| Simon Tully | National Air Traffic Services |

This publication is supported and funded by the Defence Science and Technology Laboratory (Dstl).

Dstl is the science inside UK defence and security. Dstl's AI and Autonomy programmes conduct research in data science, artificial intelligence (AI), autonomous systems and human systems integration. This work supports machine learning (ML), AI algorithm analysis and subsequent software development within the defence context, collaborating to address issues such as ethical, legal, certification, regulation and trust. For more information on Dstl's world-class expertise, please visit: https://www.gov.uk/government/organisations/defence-science-and-technology-laboratory/about

This publication is sponsored by the UK Ministry of Defence and is released for informational purposes only. Its contents should not be interpreted as representing the views of the UK Ministry of Defence, nor should it be assumed that they reflect any current or future UK Ministry of Defence policy. The information contained in this document cannot supersede any statutory or contractual requirements or liabilities and is offered without prejudice or commitment.

# Executive summary

The increasing use of artificial intelligence (AI) in many computing systems to provide functionality above that explicitly programmed has enabled improved solutions in multiple areas, including AI powered search engines, chat bots and photo manipulation software, in addition to more significant applications, such as cancer detection, and uncovering human trafficking and fraud. However, where the decisions made by such systems could greatly affect people's lives, issues such as bias, explainability, fairness, legislative compliance, ethics and trust come into play. Moreover, where such systems may be used in safety-related applications, yet more scrutiny is required to ensure that their use can be justified, and the risks understood and mitigated.

Traditional systems design and development guidance usually comprises a number of key activities that are required to produce a software and/or hardware system. These include capturing the system requirements, producing a specification, design, implementation, testing, and verification and validation (V&V). Alongside these, security considerations must be addressed, and for systems relating to safety, hazard analysis and risk assessment activities are required. Safety arguments traditionally rely on proving that the system operates in a deterministic way and that this operation is safe across all functionality.

With systems that utilise AI, these components are not built in the same way as software and hardware; therefore, the traditional systems development practices, and associated methodologies, do not account for the additional challenges faced during the development of AI systems. Moreover, the methods and techniques used to implement an AI algorithm are very different to those recommended under conventional safety standards. Hence, there is a 'gap' in being able to demonstrate that AI complies with conventional good practice and this gap will need to be addressed through alternative safety arguments.

AI components require additional activities at the design, implementation, testing and V&V stages. Consideration also needs to be given to the risk of AI adoption, the relevant governance, and ethics, as well as AI development activities such as data collection, pre-processing, machine learning and training. Aside from considerations relating to building the system, the behaviour of the system needs to be understood to the level that there is confidence in the system output. This needs to align with the human factors that influence (and are influenced by) such systems. The maintenance and operation of the system (since such systems are often a pairing of machine and user) – in particular the responsibilities of the system's owners and users – need to be clearly articulated and understood. It is critical that the system is only used in a way that is safe, and that it is only used for environments and contexts in which it has been validated.

This publication takes a safety-related view and considers the risks, challenges and potential solutions for using AI in safety-related systems. It introduces 10 key 'pillars of assurance' to underpin an assurance case for AI. All pillars relate to both the system and its components, and so should be considered together. Additionally, an effective functional safety assurance case needs to consider the interaction between the pillars and their integration into the overall system assurance case. Given the relatively short time in which AI systems have been introduced into real world use, no such consensus is yet available to deal with the different phases of the AI lifecycle. The advice herein goes alongside that provided by standards such as IEC 61508 and its various sector implementations.

Whilst the publication aims to inform and support decision making in the use of AI in safety-related applications, since this topic has yet to mature, it can be expected to evolve as knowledge and consensus grow. Therefore, following this guidance alone may not be sufficient to fully comply with the law.

# Section 1

## Introduction

### 1.1 Purpose

The purpose of this publication is to provide organisations and practitioners working in the fields of artificial intelligence (AI), including machine learning (ML), and functional safety, with information to support decision-making regarding the use of these innovative technologies in safety-related applications. This publication aims to introduce some of the risks associated with the use of AI and ML in safety-related applications and identifies some of the considerations that need to be addressed during their development and subsequent deployment.

The guidance provided here is intended to complement, but not replace, the relevant regulations and, as appropriate, international standards associated with functional safety. The use of AI in functional safety is unfamiliar territory in safety engineering and all reasonable efforts must be made to ensure that any use of AI in a safety-related application can be justified against the requirements and principles set out in the relevant regulations and international standards, albeit achieved through alternative techniques and methods.

### 1.2 What is AI?

In this publication, in line with the UK Government's National AI Strategy, 2021, we refer to AI as "Machines that perform tasks normally requiring human intelligence, especially when the machines learn from data how to do those tasks." [Ref 1]. The term AI can embrace a range of techniques, including ML, deep learning, neural networks and genetic algorithms that use data to 'learn', analyse and draw inferences. This ability to learn can range from relatively simplistic and specific behaviour for the approximation of mathematical functions (narrow AI), to the replication of human intelligence (artificial general intelligence or AGI)), to machines that are self-aware and capable of abstract and interpretive reasoning more powerful than that of humans (super intelligence).

In the last few years, the field of AI has made rapid progress, enabled by the extensive computing power that is now available. Today, AI is used in everyday applications, including search engines, facial recognition and language parsing, as well as detecting cancer [Ref 2], identifying human trafficking [Ref 3], cyber-attacks [Ref 4], fraud [Ref 5] and detection of wear in mechanical devices. In the last two years, public awareness of the use of AI has grown significantly, especially with the launch of OpenAI's ChatGPT™, Microsoft's AI-powered search engine Bing™ and Google's generative AI chatbot Bard™.

### 1.3 Safety challenges

In safety applications, any technology is expected to comply with relevant regulatory requirements. Within the UK, the enforcing authorities require that, as necessary, a system (or product) should be demonstrated to be safe. For well-established technologies, this would normally be undertaken through conformity with standards and other recognised codes of good practice, although in their absence a 'first principles approach' can be taken. Where consequences can be severe, and where uncertainty exists, it is recommended that a precautionary approach be adopted.

In this respect, the use of AI is not without problems. Instead of the conventional top-down development lifecycle, it tends to rely on the quality and completeness of data from the application domain to undertake learning and inference, as well as the use of statistical techniques to analyse the data and extract relevant parameters. An AI algorithm inherits the characteristics and imperfections of the input data, as well as assumptions underlying the statistical techniques. This approach may not produce algorithms that are understandable/intuitive to human problem solving and, as such, it may be difficult to explain the AI decision making process. Further, current implementations of AI have yet to display the same level of creativity as humans. Since its conclusions are based generally only on the observed input data, there may not be sufficient general and situational awareness for decisions impacting ethical, belief and value judgements. Finally, the methods and techniques used to implement an AI algorithm are very different to those recommended under conventional safety standards. Hence, there is a gap in being able to demonstrate that AI complies with conventional good practice and this gap will need to be addressed through alternative safety arguments. Whilst there are no generally agreed norms for 'good safety practice' for AI, this publication highlights some of the issues for consideration.

In addition, the introduction of AI will increase the overall level of complexity of a system that may be used in safety applications. This can then have an adverse impact upon its integrity, especially in circumstances where modifications and changes become necessary during development, or whilst in service. The adoption of a systems engineering approach is recognised to be an effective method for resolving challenges associated with system complexity and this can be relevant to the application of AI in functional safety. In practice, this involves the use of configuration management and control principles throughout the overall system lifecycle. These are viewed as essential measures to ensure that key requirements for safety, security and operability can be satisfied. Adherence to these principles in development and subsequent operational stages can provide an effective contribution towards consistency and continuity in satisfying safety and security objectives. In this context, configuration management and control should, as necessary, be applied to hardware, software, data, AI algorithm development, system communication capabilities, operating systems, third party libraries/code, open-source code and AI application tools, to ensure adequate traceability to assigned functions within safety, security and operational specifications.

## 1.4 AI safety opportunities

Undoubtedly, the use of AI can enable automation of applications never previously achieved. Where the consequences to safety of an inadequate decision are low, for example, in machinery applications which can be physically separated by engineering means (for example, guarded) from those it may harm, or even where the use of AI enables detection of significant threats to safety not previously detectable, the potential for an occasional 'wrong' decision may well significantly outweigh the risks. However, this may not be the case in high-consequence scenarios, especially where a body of experience has grown up in the norms to be used for digital technology to achieve a tolerable level of risk. This publication therefore addresses the potential use of AI in functional safety, specifically the use of automated systems that perform executive actions to ensure the safety and well-being of individuals, as well as the protection of assets and the environment. Some examples of such systems are:

**(a)** **Aviation:** Aircraft control systems, collision avoidance systems and flight data monitoring systems.
**(b)** **Healthcare:** Medical devices such as pacemakers, defibrillators and infusion pumps, as well as electronic health record systems and patient monitoring systems.
**(c)** **Automotive:** Anti-lock braking systems (ABS), electronic stability control (ESC), airbag systems and autonomous driving systems.
**(d)** **Manufacturing:** Safety interlocks, emergency stop systems, machine guarding systems and safety protocols for industrial processes.
**(e)** **Energy:** Safety-related systems in nuclear power plants, oil and gas refineries and renewable energy installations, decision support systems to assist in the prevention of accidents and management of hazardous situations.
**(f)** **Transportation:** Train control systems, signalling systems, railway crossing protection and traffic management systems.

**(g)** Certain types of decision support system, such as clinical decision support, management of emergency response (for example, ambulance service), management of critical alarms and situational awareness.

These systems typically rely on sensors, actuators, control systems and software algorithms to make decisions on, and respond to, unsafe situations. By putting safety measures, protocols and mechanisms into place, these systems can prevent or mitigate accidents, injuries, or damage. Although the use of AI may not provide flawless decision making, nevertheless, there are situations where it may provide more reliable and repeatable decision making when compared to existing solutions, for example, where human decision-making is prone to error. It may also be appropriate to make better decisions and act more effectively to protect people and the environment where other means of protection can be used to guarantee that the system will be sufficiently safe.

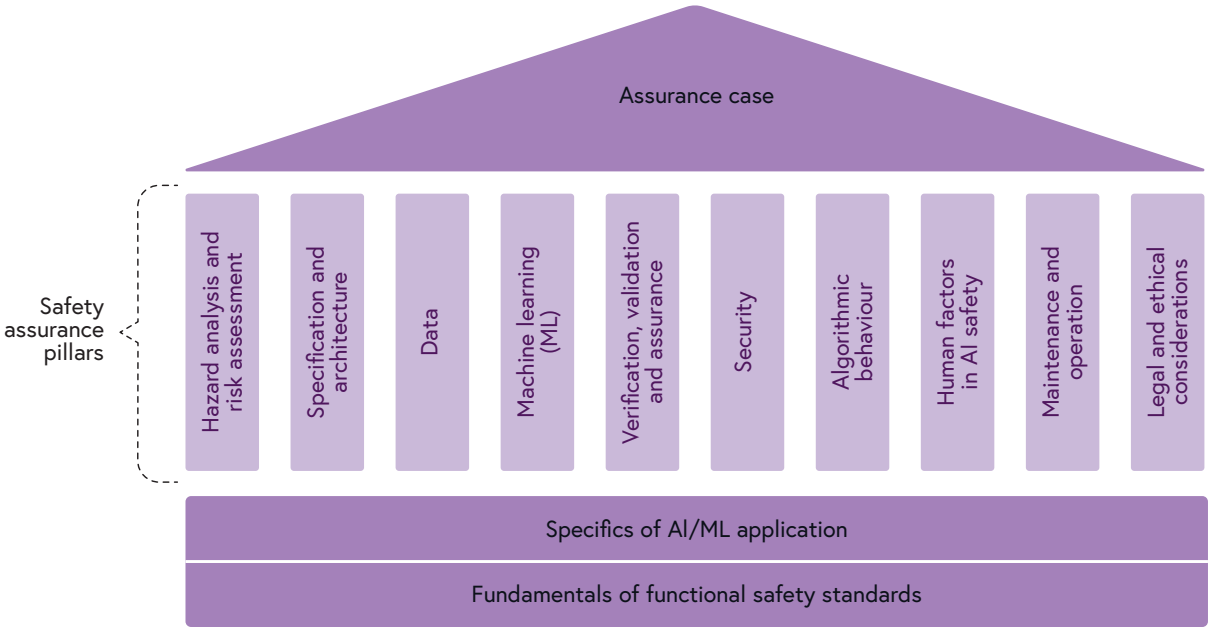
## 1.5 Assurance pillars - structure of this publication

This publication introduces 10 key 'pillars of assurance' (Figure 1.1) to discuss the considerations that need to be given to the use of AI in safety-related systems, recognising that functional safety is delivered by the overall system (i.e., sensor through to actuator including hardware, control logic, software algorithms, etc). It also lists some of the considerations that need to be given to building an assurance case for the use of AI in safety-related applications. In this context, it is likely to be easier to justify the functional safety of a relatively simple safety-related system - and the addition of AI might increase the levels of complexity and uncertainty in developing a robust substantiation of safety performance.

All pillars should be considered in relation to both the system and its AI components, although depending on the application, the pillars may not be of equal importance. Further, an effective functional safety assurance case needs to ensure that the interaction between the pillars is addressed and their integration and ranking within the overall system assurance case should be demonstrated. For example, there may be conflicting constraints when satisfying the different pillars and any resolution may affect the assurance of pillars initially considered less critical/resolved.

The pillars provided in this publication should be considered as complementary to, rather than replacing, the relevant regulations and international standards associated with functional safety (see Section 1.1). The pillars do, however, cover aspects of a functional safety assurance case that are not always fully addressed by international standards. In this sense, they have an important complementary role helping to provide confidence in the potential use of AI in achieving functional safety to support its justification. Additionally, these pillars may be replaced or augmented by additional criteria based on the continued development of AI technologies that are suitable for use in safety-related systems. Such uses may include decision support tools, or an element of the safety-related system itself. The introduction of any additional pillars and their underpinning criteria will need to take account of the greater knowledge and experience gained through practical applications and the results of any further research into specific areas (for example, validation of AI robustness in decision making).

**Figure 1.1** Key pillars for evaluating the risk of using AI for safety-related applications




The alignment between these pillars and the overall safety lifecycle model is shown in Figure 1.2. This describes the interaction in a simplified form between the overall safety lifecycle model and the pillars as additional points that require consideration when implementing AI elements within safety-related systems.

In terms of the alignment shown in Figure 1.2 it should be noted that:

**(a)** There may be overlap in the application of assurance pillars (for example, Hazard analysis and risk assessment, Specification and architecture) between phases and, in some cases, the assurance pillars (for example, Security, Verification) should be considered as transverse across all phases of the overall safety lifecycle. This is likely to be the case for Security, which is increasingly considered at the start of the overall safety lifecycle in accordance with Section 7. Likewise, although system safety architecture is not in itself a pillar for safety risks, the use of defensive architecture and AI 'safeguarding' is covered in Hazard analysis and risk assessment, Specification and architecture, and the Verification, validation and assurance pillars, respectively.

**(b)** There can be an overlap between specification and both data and machine learning, dependent upon the level of detail provided in the specification.

**(c)** It is necessary to reconsider assurance 'pillars' at the appropriate lifecycle phase where modifications occur during the development of safety-related systems, or where upgrades are required during operational phases of the overall safety lifecycle. An impact analysis should be carried out to determine the impact of the proposed modification or upgrade activity on functional safety in accordance with IEC 61508 [Ref 6].

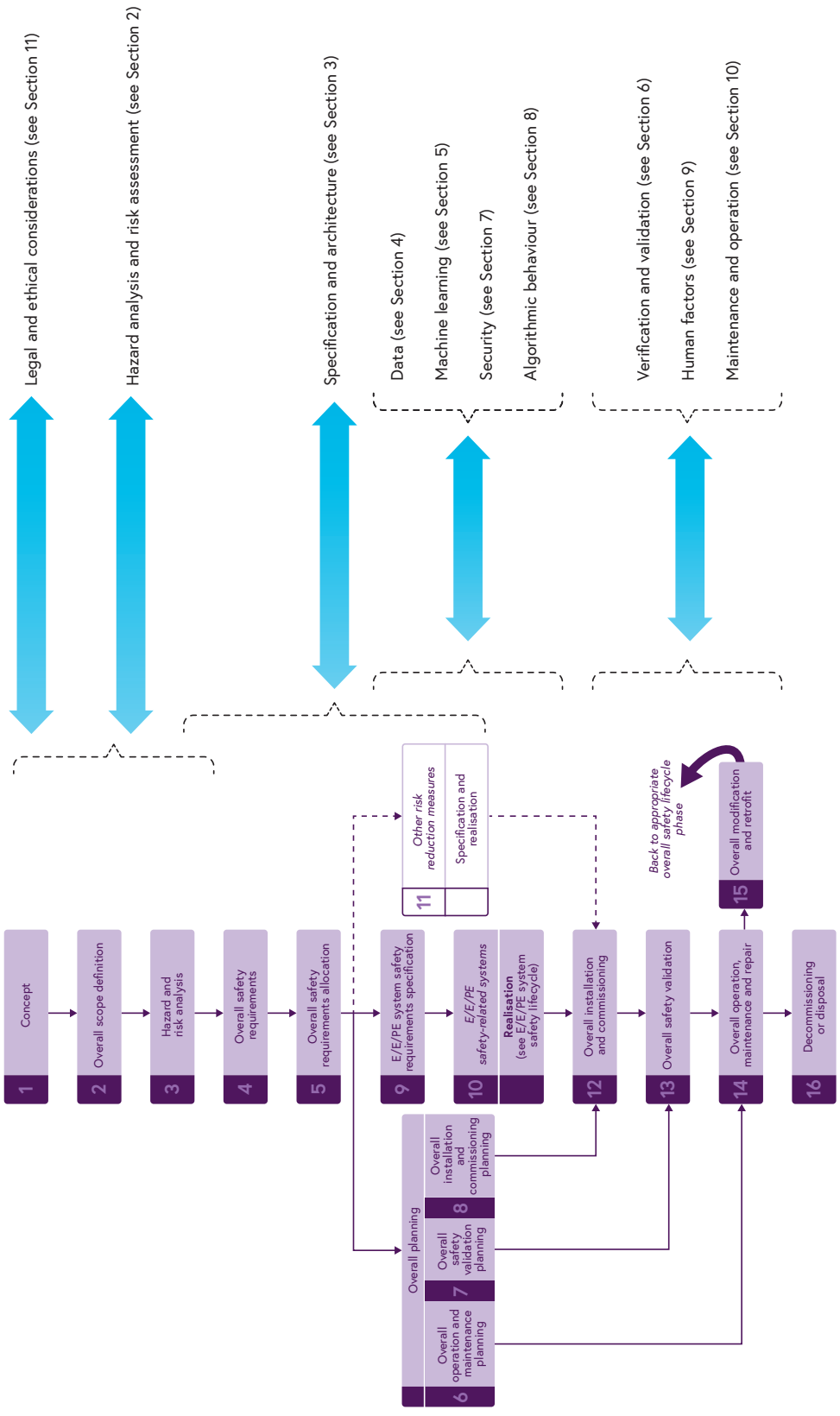



**Figure 1.2** Alignment of pillars of assurance to phases in the overall safety lifecycle model

## 1.6 AI safety - codes of good practice

Safety-related applications are generally controlled by international and national regulations, which in turn are supported by specific standards to ensure their safe and dependable operation. It should be recognised that different approaches can be applied in different industry sectors to fulfil principles of functional safety set out in IEC 61508 [Ref 6]. The relevant international standards, include:

**(a)** IEC 61508, the basic functional safety publication from the IEC [Ref 6].
**(b)** ISO 26262, the ISO standard on road vehicles [Ref 7].
**(c)** BS EN 50126/IEC 62278, 'RAMS', IEC 62279, 'software', and BS EN 50129/IEC 62425, 'signalling hardware', for railway applications [Refs 8, 9 and 10].
**(d)** DO-178C, *Software Considerations in Airborne Systems and Equipment Certification*, the primary document by which certification authorities approve commercial software-based aerospace systems. The document is published by RTCA [Ref 11].
**(e)** Other sector specific functional safety standards, for example, those commonly used for machinery (IEC 62061 [Ref 12] and ISO 13849-1 [Ref 13]) or process industries (IEC 61511 [Ref 14]).

In practice, sector specific standards provide guidelines for development, verification and validation, operation and maintenance of systems to ensure that safety objectives are identified and met. The specific processes advocated by these standards do not currently address solutions generated by AI and, indeed, in the past have deterred its application, although guidance is now under development in a number of ISO, IEC and other communities, including:

**(a)** ISO/IEC TR 5469:2024 *Artificial intelligence - Functional safety and AI systems* [Ref 15], published in January 2024; this document describes the properties, related risk factors, available methods and processes relating to:

  **(i)** use of AI inside a safety-related function to realise the functionality;
  **(ii)** use of non-AI safety-related functions to ensure safety for an AI controlled equipment; and
  **(iii)** use of AI systems to design and develop safety-related functions.

**(b)** PAS 8800 [Ref 16], nearing final drafting for automotive applications;
**(c)** ISO/IEC AWI TS 22440 *Artificial intelligence - Functional safety and AI systems - Requirements* [Ref 17], approved for development by ISO and IEC in September 2023 to translate [Ref 15] into preliminary requirements for standardisation of AI used in safety applications. (The planned publication date is in 2026); and
**(d)** International Atomic Energy Agency [Ref 18].

Whilst guidance on the application of AI in functional safety is under development, this is at an early stage of maturity and at the time of writing, there is no generally agreed body of knowledge as to what constitutes a sufficient process for assurance of its safety.

# Section 2

## Hazard analysis and risk assessment

### 2.1 Definition of 'hazard' and 'risk'

A hazard to safety is defined in ISO/IEC Guide 51 [Ref 19] as a "potential source of harm", with the clarification that this includes danger to persons, arising within a short time scale (for example, fire and explosion), and also those that have a long-term effect on a person's health (for example, release of a toxic substance). A hazardous event is defined in IEC 61508 as an "event that may result in harm". Note that the term 'risk' in the safety discipline is defined as "the combination of the probability of occurrence of harm and severity of the harm" [Ref 19], whereas the ISO definition [Ref 20] is more broadly defined as the "effect of uncertainty on objectives". In the context of this publication, the ISO/IEC Guide 51 definition is used unless explicitly stated otherwise.

In terms of current practice in accordance with IEC 61508, although AI components may not be used to directly implement safety-related applications, the impact of using AI on safety must still be considered. For example, if a sensor network uses AI-based pattern recognition, a failure in the sensor network could impact a functional safety component that uses the sensor input. It should also be considered that the output of an AI component could be to another machine or autonomous device which might in turn influence the safe behaviour. Further, the AI system might itself increase the number of demands to which a safety-related system must respond, thereby increasing the risk that it may respond incorrectly, or place the system in an abnormal recovery state that has an increased risk of harm.

### 2.2 Conducting a hazard analysis and risk assessment

Multiple different techniques are used to identify hazards and assess the risk associated with a hazard from concept throughout the lifecycle of conventional safety-related system development, including during operations and maintenance. The techniques range from simple lists produced from previous experience/brainstorming, through to systematic analysis techniques to examine potential deviations from expected behaviour, depending on the complexity and criticality of the application. Common techniques include Hazard and Operability Studies (HAZOPS), Functional Failure Analysis (FFA), Failure Mode Effect and Criticality Analysis (FMECA), and Fault Tree Analysis (FTA). Comprehensive descriptions of each of these techniques can be found in, for example, [Refs 21 and 6]. Techniques from software systems dependability analysis, such as Systemic Cause Analysis [Ref 22] or System Theoretic Process Analysis [Ref 23], which focus on analysis of hazardous events and incidents at a systems level, may also be useful.

As with other complex systems, a factor when analysing risks arising from a system incorporating AI components is that the behaviour of the system may depend not only on its individual parts, but also on the interaction between the parts (emergent behaviour). This can lead to accident sequences not identified during the formal analysis of the system. Whichever type of analysis technique is being used, it's important that potential causes of failure from an AI component are understood and considered during hazard analysis and risk assessment. At the functional level, a set of guidewords is typically used to prompt investigation of deviations from the intended function:

**(a)** none;
**(b)** too much, too little;
**(c)** too early, too late;
**(d)** as well as;
**(e)** part of;
**(f)** opposite; and
**(g)** other than.

At the component level, it may also be appropriate to consider the types of generic failure event that are specific to AI and data science and whether they could lead to hazardous situations such as:

**(a)** inaccurate prediction;
**(b)** distribution shift;
**(c)** lack of fairness and equality; and
**(d)** model uncertainty.

A further consideration is the potential for risk due to hazards resulting from functional deficiencies of the intended functionality. Often termed the Safety Of The Intended Function (SOTIF), this is further addressed in [Ref 24].

The failure modes that that could arise from an AI component will be dependent on the AI technology used. It's important to consider the technology, regulatory environment and operational context when considering risk. Currently, the dominant ML method is deep learning which uses multi-layered neural networks. The deep learning method makes visibility and validation of generated models particularly difficult. Examples of high-level failure modes of such a system include:

**(a)** Model input error, such as a sensor error or an input outside the training data set of the model.
**(b)** Incorrect output caused by model bias, for example, where a model has given undue weighting to a condition that occurs frequently in a particular data set, but which is not representative of a more general case. For example, if training data only included humans from one ethnicity, it may fail to identify those from other ethnic backgrounds.
**(c)** Incorrect output caused by lack of algorithm robustness, i.e., the extent to which an AI algorithm is able to maintain its performance under any circumstances, including unexpected inputs, external interference or harsh environmental conditions.

In practice, it may be necessary to augment AI components with conventional safety measures to avoid any adverse effects arising from incorrect outputs due to a lack of robustness. This defensive architectural approach is commonly used to manage risks in safety-related systems that do not incorporate AI technology and, as such, it is important that the application of AI in these applications justifies how it is integrated within the overall design of the system so as not to undermine safety integrity.

In any hazard analysis or risk assessment involving safety-related systems that include components with AI, the inclusion of a data scientist as a participant could help identify failure modes that stakeholders may not be familiar with. AI-based discovery tools have also been developed which could assist in prompting and extending the collective mental model of the team conducting a hazard analysis or risk assessment.

A benefit of generic models trained and tested on large amounts of data by an open-source community, is that they are likely to be less sensitive to variation in training data. Emerging techniques include working with small data sets to fine tune a generic model. The model can be tuned with curated content for very specific use cases using smaller and more controllable sets of incremental training data. In general, the use of proven AI design patterns and components as a means of simplifying design and development should be encouraged as this could make risk management more practicable.

## 2.3 Maintenance of the hazard and risk assessment during change

The management of AI components throughout their lifecycle is critical in their deployment. The behaviour of AI components can change over time if the component is enabled to learn from observed data, or if the operational characteristics of the system change. An analogy can be drawn with task-based risk assessment where the risk of a previously assessed work task may vary according to the site conditions at the time the work is carried out. This can be addressed by a process of 'dynamic risk assessment'. For example, a 'toolbox talk' may be delivered immediately before work commences to provide an update on the current situation. While changes in response to data input to the AI component can

be controlled, a similar process of dynamic risk assessment, aligned with regular monitoring, feedback and adjustment of the AI component, should be considered to review changes in risk associated with the component. Governance tools for AI are available and their use may be helpful in this context.

Frequent testing, including impact analysis, of changes to a model can be facilitated using automated test routines for regression test against original pass criteria. For applications where AI acts as a digital assistant (performing tasks that would require a human to be competency tested, for example), there may be a case for defining competency criteria for the digital assistant. This could be based on requirements such as accuracy and precision and could be tested at regular intervals.

## 2.4 AI safeguarding

To protect against unexpected behaviour, either as a result of errors during the initial development, or as a result of change during operation, an overall layer of governance and safeguarding is likely to be required - effectively acting as a safety jacket. In many cases, a deterministic rule-based protective layer will be required to prevent the AI component executing behaviour outside of safe limits. For example, an AI assistant performing an automated approval function may not be allowed to approve safety-related activities without human oversight (see also Section 9). Governance may also include regular audit of the model to protect against change.

## 2.5 Summary and conclusions

This section has introduced the concept of hazards to human safety (both short and long term), the risk of occurrence of such a hazard, sources of behavioural anomaly that could result in an increased risk to human safety, and some of the methods that have been used to identify such risks. The ways in which AI systems in particular may increase such risks are also discussed.

## 2.6 Considerations to control the impact of the use of AI on safety-related systems

Care needs to be taken to control the impact of the use of AI and, as such, it is necessary to consider a range of factors in relation to potential hazards and risks that may arise, including:

**(a)** The framework for use of different types of AI technology in [Ref 15] should be used as a guide as to the potential impact on safety risks; in particular, this report identifies a framework based on the degree to which the application can directly affect safety, and the extent to which the technology used can be justified against principles required to assure safety within that application. The baseline for the framework is the underlying principles in IEC 61508.

**(b)** It is important that the impact of the AI on the overall safety-related system, and hence, the way in which it could impact the harm arising from the use of the system, is understood. There are many hazard analysis and risk assessment techniques which are available to assess such impact at the functional level, as well as through consideration of the ways in which errors in the AI algorithm or its underlying data can lead to a hazardous condition. Appropriate consideration should be given to both the source of failure and its impact on a hazard - unless it can be shown that there are no conditions under which the AI can impact human safety.

**(c)** Analysis techniques need to be applied that can consider potential hazards arising from sources internal to a safety-related system that might contain AI, including hardware failures, communication errors, integrity of operating systems, the use of third-party libraries/code, open-source code and AI application tools (for example, those used in design, configuration and maintenance). This focus on hazards arising from failure of the AI is important and complements the AI 'safeguarding' outlined in Section 3.4 where features, such as diagnostic capabilities and other measures to detect and respond to failures (for example, watchdogs), can be used to improve safety integrity in accordance with IEC 61508.

**(d)** Ability to avoid harm depends not only on individual components/systems, but also on how, as a whole (including linked systems, human activity and the environment), these interact together; this should be taken into account. Further information can be found in the Human factors pillar, which considers the interactions between humans and built systems, as well as understanding their limitations.

**(e)** Provision of a mechanism to ensure the output cannot exceed the safe boundary of the operation.

**(f)** The nature of AI technology is such that it is difficult for a human to understand the underlying rational for its behaviour, hence the need for point (e).

Notwithstanding the above, by enabling a measure of control over phenomena not previously controllable, AI could potentially improve safety. However, it is questionable whether it could be sufficiently trusted as a replacement to conventional safety-related systems. Its potential to increase a demand on a conventional safety-related system should also be considered.

# Section 3

## Specification and architecture

### 3.1 Derivation of the safety requirement

A specification of a safety-related system can be considered in two parts: what does the user need and how will the system behave. Ideally the two will be equivalent.

In terms of the user need, there is little difference between a conventional system and one based on AI. The user will foresee a benefit in deploying the system, but that may introduce or increase the risk of harm. Therefore, a 'needs' specification will be in two parts; definition of what the user requires of the system to obtain the benefit, and how to control the risk of harm to acceptable levels. The 'user requirement' specification is independent of the technology to be used and therefore conventional functional safety approaches should be used [for example, Refs 7 to 11 and 14 and/or other appropriate sector specific standards].

In conventional functional safety standards such as [Ref 6], the safety specification defines the safety functions (the functions that are intended to achieve or maintain a safe state) and their safety integrity levels (SILs), the confidence that the safety-related system will perform the required function(s). The safety functions are derived by a systematic consideration of the hazards associated with the use of the system and the functions that are necessary to control them. The safety integrity level(s) relates to the confidence that each safety function(s) will be effective and is based on a mapping between levels of risk and engineering judgement as to the methods and techniques to apply to control the risk.

### 3.2 Definition of the actual behaviour

The difficulty comes once decisions are made on how the functionality will be apportioned, both logically and physically, between different systems and equipment (the architecture) and the choice of the technology to be used for each. For safety-related systems, the ideal technology is fully predictable under all conditions and will 'fail safe', i.e., where injury to humans (short or long term) cannot occur. In practice, absolute adherence to these constraints cannot deliver the benefits society expects. For example, air travel would be impossible, power systems could not be efficiently delivered and computer-based medical devices could not operate.

Thus, in practice, use of computer technologies in safety-related systems is widespread and conventional functional safety standards have been developed to minimise complexity (thus maximising predictability of behaviour), encourage methods that are less likely to introduce error during development and provide alternative means of protection should residual errors in behaviour exist.

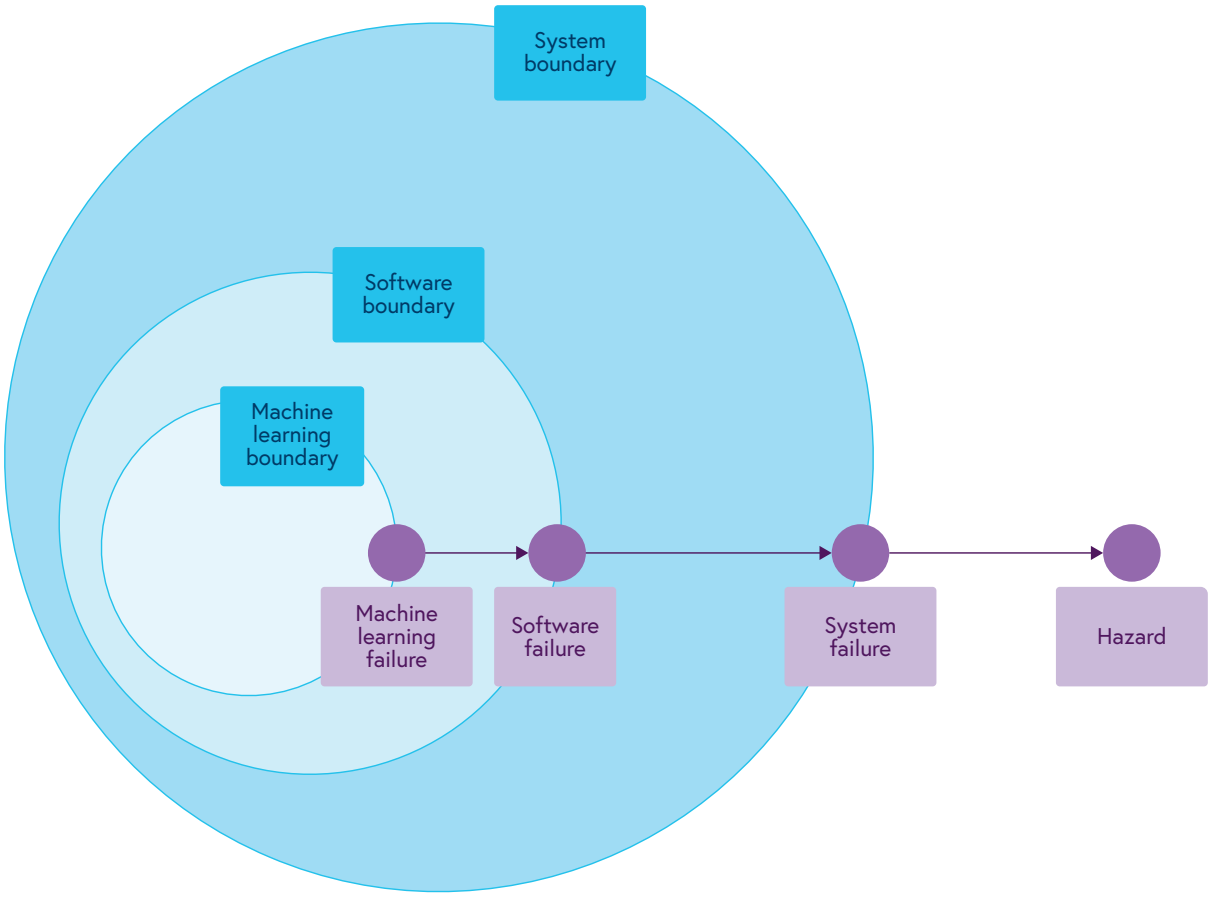
Unfortunately, as shown in subsequent sections of this guidance, the methods and techniques advocated in functional safety standards do not comprehensively address the types of error associated with AI components. Although some aspects of AI components can adhere to traditional requirements, for example, the learning algorithm of a neural network can be assured against conventional standards. The difference with functional safety standards can be addressed, in practice, by constraining the impact of AI components on system performance, using a combination of physical guarding, interlocking schemes or other engineering measures (see Section 3.4) to avoid or control risks that may arise from incorrect decision making.

There is also the question as to how the safety functions are apportioned between components. In conventional safety-related systems, 'platforms' (such as a programmable logic controller or database servicers/services) will be selected and the application algorithms will be 'programmed' onto these platforms. Multiple platforms may be selected to provide diversity, redundancy and segregation of different levels of safety criticality. In addition, diagnostic functions may be identified to detect and react to faults and failures in the primary safety functions. The architecture will be constructed with traceability through to the safety functions, such that the multiple layers of protection can be evaluated to assure the required safety integrity level (SIL) as well as to ensure that each sub-function does not interfere with the operation of the whole. Traceability of the AI components and their architecture to the top-level functional requirements may not be straightforward.

## 3.3 Behavioural uncertainties

It is important that the contribution made by the behavioural uncertainties associated with the AI components to the hazards be articulated in the form of failure conditions that can lead to its occurrence in a given application. Whilst this allows the AI component contributions to the potential hazard to be better understood within the context of the safety-related system and its architecture, appropriate measures need to be specified to manage the risk of occurrence of failure conditions that can lead directly or indirectly to a hazard. Hawkins et al. [Ref 25] propose a modification of the AMLAS [Ref 26] approach, to illustrate this (see Figure 3.1).

**Figure 3.1** Simplified chain of failure events



# Section 3 – Specification and architecture

Whilst it is important that safety requirements are specified in an unambiguous and complete manner, it may not always be possible to define the safety functions in such objective terms, even in a conventional safety-related system. A functional safety specification can be considered as a set of proxy requirements of the human's intent for the system behaviour. For example, in the case of a 'simple' traffic light system, the safety function might be defined as "A vehicle shall not pass a red traffic light". At first glance, this statement may appear unambiguous. However, when considered in greater detail, it becomes clear that, in real-life, there are many scenarios in which a vehicle may be required to pass a red light, for example, when it becomes clear the traffic signal is not working or if instructed by a police officer. Unless the specifier has anticipated the complete set of use cases, even a conventional safety-related system may behave in an unexpected way.

In a conventional safety-related system, even if incompletely defined, there is a traceable link between the observed behaviour and the safety function/integrity level required to meet the user expressed need. In the case of AI, this can be a challenge since rules may be generated which do not accord closely with human representations of the problem domain. Further, outputs are derived probabilistically and may not react in an expected manner to every possible environment state. It may be difficult to identify a missing use case or to explain why the system has behaved as it has.

To avoid incorrect specification, it may be necessary to take account of the proposed implementation technology during specification of the high-level requirement. This approach would require previously unknown low-level events to be linked to the top-level hazards to ensure that overall system safety claims can be sustained.

Although specification of AI requirements and behaviour is a developing field, Heyn et al [Ref 27] have articulated the potential challenges across four problem areas: contextual definitions and requirements; data attributes and requirements; performance definition and monitoring; and human factors. These are overlaid on system capabilities, for example, functional behaviour, safety and cyber resilience, etc. This framework could be a useful tool for developing requirements that meet the intent of the safety-related system. To increase confidence, it may be prudent to validate the specification with all relevant stakeholders associated with delivery, assurance, use and maintenance of a safety-related system that incorporates either directly or indirectly (for example, digital assistant) AI technology.

## 3.4 Summary and conclusions

This section has provided a brief overview of the considerations when developing safety requirements specification for AI systems. At the level of specification of the user need, the approach should adhere to that used for any safety-related system, as the 'need' should be agnostic to the means of implementation. Further information can be found in functional safety standards, for example, IEC 61508.

However, once a decision on technology has been made, the detail in the specification will need to take account of the characteristics of the development technology. In the case of AI for example, the specification will need to ensure that the training set does incorporate the constraints necessary to ensure safe behaviour and to focus the specification on the required outcomes under all foreseeable conditions, including the impact of erroneous AI decision making.

## 3.5 Considerations for good practice during specification of AI systems in safety applications

The role of specification and architecture is important in the context of understanding requirements for safety-related systems and the uncertainties associated with AI. In practice, it is necessary to consider a range of factors for this pillar, including:

**(a)** Given the cost and general lack of available data for some applications, it may be tempting to develop the specification of the user need from the available data. This should be avoided since high-level requirements should be implementation agnostic. Instead, conventional functional safety techniques based on consideration of the hazards and potential accident scenarios should be used to develop the highest-level specifications.

**(b)** Using the requirements to drive the collection and curation of data should help ensure that it meets the intended outcome. There can be significant investment in training data and time just to evaluate an early concept.

**(c)** Attention needs to be given to address the behavioural uncertainties associated with AI to establish the measures and techniques that should be applied to determine failure conditions that can lead to hazards in any given application in functional safety.

**(d)** Ensure that the context of operation is clearly articulated. For specifications derived through ML, it is important that the data used is drawn from a sufficiently broad range of scenarios that might occur in the operation of the system.

**(e)** At the same time, it is not simply a case of more data is better. ML components may over-train on expected scenarios and be unable to generalise behaviour when 'edge cases' arise. It can be these abnormal conditions that give rise to accident sequences.

# Section 4

## Data

### 4.1 Data types

The categorisation and use of data in safety-related systems that incorporate AI is very different to that of a conventional safety-related system and the quality and completeness of the data sets has the potential to be a significant source of systematic failures. Additionally, the quality of data can impact security.

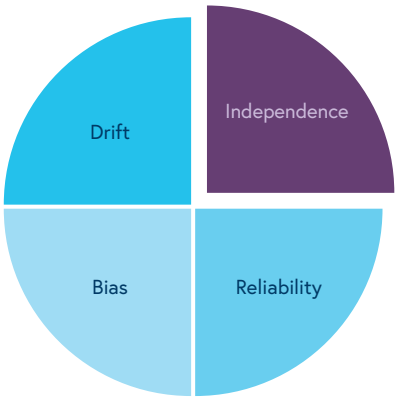
Data used within an AI system is normally divided into the following four categories:

1. **Input data:** Data produced by the sensor elements of the safety-related system that is used by the AI model to make decisions.
2. **Training data:** Specific data sets used to teach the AI models or algorithms to make the proper and correct decisions.
3. **Test data:** Data set specifically used to test, verify and validate the AI model.
4. **Experience data:** Data collected during the operational phase.

### 4.2 Avoiding systematic errors due to data

When specifying and compiling the data, there are a number of factors that should be considered in order to limit the introduction of systematic errors. These factors include:

**Figure 4.1** Factors that can be used to limit the introduction of systematic errors




**(a) Independence:** There should be sufficient independence between training and test data sets to ensure a robust testing and validation of the safety-related system. Without independence it will be difficult to detect when the AI model is unable to generalise to the input data, known as 'overfitting'. When this happens, the AI model may not perform the required safety functions in the real-world and to avoid the onset of a hazard, it may be necessary to demonstrate that effective measures are provided to address this risk. Independence can be built into the development process and could be linked to the level of safety mitigation required. Independence could be further enhanced by using separate teams for development and testing.

A consideration in developing the safety assurance case is that current functional safety standards lay out the minimum random and systematic failure rates for safety-related systems; it is unlikely that even these failure rates can be demonstrated purely by testing the AI system, and therefore additional forms of safety argument will be required (for example, based on analytical methods).

**(b) Reliability:** The input data provided to the AI model needs to be reliable and have an integrity commensurate with the level of safety risk mitigation required. Random or systematic failures within the input sensor sub-system could provide erroneous data to the AI model. This may lead to incorrect decision-making and ultimately a failure of the safety-related system. In essence, safety-related systems can suffer from similar sources of failure and hence, traditional functional safety standards, such as IEC 61508, can be used to design the subsystem elements (for example, sensors) and ensure they achieve the required level of safety integrity.

**(c) Bias:** If the training and/or test data sets are not sufficiently diverse or representative and/or do not cover the entire problem 'space', the AI model may develop unwanted bias. This may lead to the decision-making algorithm allocating inappropriate 'weights' to the control parameters to be used in decision-making, or even to the complete omission of parameters needed to cover rare but foreseeable conditions. In this case, the decision-making algorithm used by the safety-related system may fail to protect against specific threats or operational scenarios. The design process should include assessment of data sets and activities with the specific aim of identifying potential bias.

**(d) Drift:** There are many reasons why input data may drift over time. This can be down to external factors, such as seasonal changes in the operating environment, or internal factors such as degradation/calibration of sensor systems. When potential data drifts are identified, the safety-related system can be designed to incorporate correction factors. These adjust for the fact that the existing data (from the model) is not representative of new data coming in from the environment. It is important to monitor data during operation of the safety-related system in order to highlight previously unanticipated sources of data drift. When this occurs, re-training, testing, and verification and validation of the AI model will be required, including updating the assurance case to demonstrate the system remains safe before it is re-introduced into operation.

Additionally, the spread of input data across the input domain space should be such that the data is complete, and the data set is balanced. Input data can fall into one of four 'domain spaces'. These can be considered using the following criteria [Ref 28]:

1. **The input domain space:** This is the set of inputs that the software implementation of the AI model can accept and is a feature of the underlying computer platform (i.e., the way data is represented in the computer hardware/software).
2. **The operational domain space:** This is the set of inputs that the model may be expected to receive when used for decision making within the intended operational domain. This can include inputs due to operational environmental effects.
3. **The failure domain space:** This is the set of inputs the model may receive if there are failures elsewhere in the system.
4. **The adversarial domain space:** The set of inputs the model may receive if it is being attacked by an adversary (for example, a cyber-attack). Even small changes achieved through malicious activity could cause the AI algorithms to reach a false decision – for example, a road sign with 'stickers' placed on it may be misclassified, or other deliberate interference may alter the system's input data.

## 4.3 Addressing the challenge of data quantity

One of the challenges faced when training ML algorithms, is the sheer quantity of data required to learn complex tasks [Ref 29]. Acquiring such data is often difficult, costly and labour intensive. Moreover, for particular domains, there may be additional challenges due to the practicalities of observing, or even predicting, 'edge' cases, i.e., when there is a change in the parameters driving the behaviour of the decision-making system. For example, if the system is trained only on normal operation, parameters that control the system during abnormal or failed conditions or at the extremes of the operation may not be present in the learned behaviour. It is essential to ensure that all foreseeable conditions are represented in the training data and that the transition points are represented to ensure that the models will be robust when deployed in the real world.

A number of learning techniques have been proposed to address the quantity of data which leverage the fact that deep neural networks seem to learn general features that are transferable, and hence, can be reused by other similar tasks [Ref 30]. The main idea of these approaches is to:

**(a)** copy parts of the pre-trained source models into the target model;
**(b)** add one or more randomly initialised (untrained) layers into the target model, such that the last layer now matches the target's label space which can minimise the possibility of human error in mislabelling the data and introducing bias or other issues; and
**(c)** train the model using labelled target domain data.

However, these techniques cannot be used if the data originates from custom sensors that are specific to the domain in question. In such cases, pre-trained models rarely exist or may be too sensitive to share.

Unsupervised domain adaptation, where data from a diverse source is used to train a model for use in the problem domain, is another transfer learning technique. This offers the potential for using synthetic data – i.e., data generated by simulation of the problem domain, or by another (for example, scientific) model that already exists. Synthetic data can provide ample data that may otherwise be difficult, impractical or costly to obtain, including the representation of edge cases thereby helping to provide complete data sets. Another benefit of simulating training data is that it eliminates the need for human analysis of different data cases: manual labelling can result in incorrect or biased data [Ref 30]. It should be noted that the generated data can only be as representative of the problem domain as that of the model from which it is derived. Consideration must be given to the trade-off between generating sufficient data quantity and maintaining diverse and representative data that will not introduce bias.

## 4.4 Summary and conclusions

This section has looked at the categorisation and use of data in safety-related systems that utilise AI. Different types of data have been described and potential issues discussed. In dealing with these issues, it is important to consider the spread of data across the input domain space, how to ensure the data is complete and balanced, and how to acquire and manage sufficient amounts of high-quality training data.

## 4.5 Considerations for good data management practices in safety-related systems

Care should be taken when specifying and compiling all categories of data used within an AI system. Specifically, Independence, Reliability, Bias and Drift should be addressed to:

**(a)** Ensure sufficient independence between training data and test data.
**(b)** Ensure the input data provided to the AI model is reliable and has sufficient integrity.
**(c)** Incorporate an assessment of data sets and activities to identify potential bias.
**(d)** Ensure that operational data is monitored to highlight previously unknown sources of data drift. Re-training, re-testing and re-verification/re-validation of the AI model may be required, including update of the assurance case.
**(e)** Overcome the challenge of acquiring sufficient quantity of data to learn complex tasks. Techniques such as transfer learning, unsupervised domain adaptation and the use of synthetic data can be beneficial.

# Section 5

## Machine learning (ML)

### 5.1 Importance of machine learning for engineering applications

Machine learning (ML), essentially the application of statistical techniques to extract 'information' from data, is useful in a range of engineering applications, for example:

- **(a)** to develop machines capable of performing tasks that imitate human intelligence, such as understanding natural language, recognising images, making decisions and solving problems;
- **(b)** to create new knowledge and discovery through the analysis of vast amounts of data, thus enabling the creation of innovative solutions to real-world problems; and
- **(c)** to automate existing processes and tools to improve efficiency, reduce human error and enhance existing capabilities.

ML is not a mandatory characteristic for apparent AI (for example, 'knowledgeable/intelligent' behaviour can also be derived from human knowledge [Ref 31], deduced from scientific theories or from computational optimisation such as genetic algorithms [Ref 29]). However, the use of ML to develop AI systems leads to a much broader range of applications, including those for which there is little 'a priori' knowledge. This leads to its own set of challenges.

There are many definitions of ML however, in the context of this publication, ML is defined [Ref 32] as the "process of optimizing model parameters through computational techniques, such that the model's[1] behaviour reflects the data or experience." ML can be used to generate a representation (or algorithm or model) of the problem space, by using statistical techniques to extract patterns and to identify classification parameters from data.

### 5.2 Types of machine learning

There are three main categories of ML: supervised learning, unsupervised learning and reinforcement learning:

1. **Supervised learning:** Where a set of pre-correlated input and output data pairs are presented to the ML algorithm.
2. **Unsupervised learning:** Where the learning algorithm identifies statistical correlations between the characteristics of the input data and output.
3. **Reinforcement learning:** Where the algorithm representing the correlation between input data and output is driven to maximise a particular 'cost' function.

**Supervised learning** [Ref 33]

With supervised learning, the characteristic signatures of the input data that distinguish individual outputs are identified. It is important that the training data covers the entire range of inputs the system will experience in operation – for example, if a model is trained to recognise a car based on images taken from the side only, it would be unlikely to correctly identify a car based on an image taken from the front or back. The process of supervised learning is human intensive as the data set must be first classified by human intervention, thus introducing the risk of defects and bias into the model as a result of human error, either in interpretation or in classification. The solution may also be sub-optimal as the model is based only on the distinguishing features identified through human interpretation, rather than having freedom to identify previously unknown relationships between input and output.

1 internal variable of a model that affects how it computes its outputs.

Supervised learning can be onerous in terms of human activity to prepare the data but can be appropriate when the distinguishing classification characteristics are known.

**Unsupervised learning** [Ref 34]

With unsupervised learning, the learning algorithm discovers for itself the distinguishing features that are used to classify the data. Unsupervised learning can identify structures within the data that are not sufficiently explicit to be apparent to human interrogation. The approach has several benefits, including a reduction in the amount of human interaction required to prepare the learning data sets, the elimination of systematic defects being introduced due to human error and the capability to classify data where no previous theories about the relationship between the input and output exist.

With unsupervised learning, the algorithm can 'cluster' the data into classes and when presented with unseen data it can allocate it to the different classes. Problems occur where the clusters are either too narrow or too broad. There is also a risk that the model may pick up on false distinguishing features within the data. There are anecdotal stories of an ML application designed to identify images of fish, which instead classified the data according to the fingers holding the fish!

**Reinforcement learning** [Ref 35]

Reinforcement learning is based around the optimisation of a 'cost' function. In simplistic terms, the learning algorithm uses pre-existing parameters to predict the desired output. The 'correctness' of its prediction is assessed against a cost function, and the parameters within the model are updated based on a reward or penalty dependent on the 'closeness' to the cost function. In this manner, Markov Decision Processes (MDPs) gradually reach the correct answer through a series of successive attempts. The next optimal action is guided by variables such as the environment, the agent's previous action and rewards.

As with an optimiser, disadvantages include speed of convergence, local minima within a multi-dimensional algorithm etc. Further, the decision on how to apply the reward can have an influence on the effectiveness of the learning. For example, an autonomous truck where the intention is to detect nearby humans, that is rewarded for each successful detection of a human, may be driven to 'look for' humans in order to be rewarded.

The choice of learning algorithm will depend on the nature of the application. For example, where the problem domain is well understood, supervised learning may be the most appropriate. Where the goal is to identify patterns in previously unknown domains, unsupervised learning may be the only viable route. If there is limited knowledge about the domain and the only information is about the required behaviour of the system, reinforcement learning may be most appropriate. The challenge for a safety-related system is to ensure that whatever the learnt model, its behaviour can at all times be demonstrated to result in 'safe' outputs.

## 5.3 Impact of data on machine learning

Every ML application is unique, with its own problem domain, level of 'a priori' knowledge and data characteristics. The choice to use an ML technique, if at all, should align with the problem requirements and available resources. The notion of one learning method being superior to another will depend on the nature of the data and the characteristics required in the problem representation/decision making model. Two particular considerations are key:

1. **Type of data:** This has a significant impact on technique selection. Types of applications include:

**(i)** the type of classification (for example, whether for image feature detection, anomaly detection);
**(ii)** the statistical assumptions (regression techniques) underlying the analysis of the relationship between the input data and the output prediction (for example, wear detection, predictive maintenance);
**(iii)** algorithms to identify 'clusters' of data points (for example, to identify 'outliers' in anomaly detection, or relationships between objects of interest); and
**(iv)** dimensionality reduction (for example, data compression, noise reduction).

2. **Amount of data available:** The method should match the amount of data to prevent overfitting or underfitting. Overfitting occurs when the model has a surfeit of data and attempts to produce an algorithm that reflects individual data points, rather than a 'best fit' model. Underfitting happens when the model has too little data and overly generalises between data points. The ideal performance occurs when the model is well-balanced, i.e., when all important parameters defining the model are distinguished, but individual data points are subsumed within the defining parameters.

Optimal learning needs to balance a range of performance considerations. For example, accuracy and speed of prediction is important but not to the detriment of robustness. Consider that apparently anomalous data may be removed during statistical learning - to avoid deviation from the 'ideal' and reduce the number of parameters that need to be considered during prediction. However, removal of outlier data can compromise a system's robustness (for example, by eliminating examples of adversarial attack, or obscure 'edge' [rarely seen data] points). The training process should continue until the model (algorithm) achieves sufficient robustness, i.e., enabling it to produce pre-determined and repeatable outputs for full coverage of the 'input space', including reasonably foreseeable misuse.

During the process of ML, data from the problem space is divided into subsets.

**Training data:** Used to fit or train the initial model and is the primary source of information for updating the model's parameters. The aim of the training process is to minimise prediction error, and hence, the quality and quantity of the training data can significantly impact model performance.

**Testing data:** Used to evaluate the accuracy of prediction of a trained model. This data is typically excluded from the training process and is used to assess the accuracy of model prediction when exposed to previously unseen data. The performance of the model can also be used to fine-tune parameters.

**Verification data:** Used to fine-tune the model's parameters. This data is used to determine the best set of parameters for the model and can help to prevent overfitting or underfitting.

In conclusion, the data used for AI learning should be diverse, representative and of high quality. The choice of data subsets and their use in the training, testing and validation processes can significantly impact the performance and accuracy of the resulting models.

## 5.4 Online vs offline learning

The means to collect data for ML is discussed in the previous section. This can be undertaken during development and during operation of an AI-based system. However, the decision on when to implement newly learnt behaviour is a key differentiator for safety-related systems.

'Static' (offline) learning permits the data to be gathered and analysed but without change to the operational system. Dynamic (online) learning would permit the control parameters to be modified whilst the system is operational to enable it to dynamically adjust to its environment or optimise its behaviour.

It should be noted that in cases where ML can impact safety, online learning (i.e., learning whilst operational) is not consistent with conventional functional safety standards, as any change requires that a comprehensive impact assessment and re-verification/re-validation activity is carried out before the safety-related system is put back into operation.

## 5.5 Summary and conclusions

This section has summarised the opportunities associated with ML, including the ability to address problems not previously amenable to human endeavour, and it has highlighted the factors affecting the adequacy of the resulting decision-making algorithms. It has also illustrated that the adequacy is highly dependent on the characteristics of the data, the coverage of the problem domain, the assumptions underlying the statistical techniques applied to the data and the extent to which the extracted parameters are both sufficiently comprehensive but not overly detailed. Further, the use of learning to modify a safety-related system still needs to be subject to appropriate assurance processes.

## 5.6 Considerations for machine learning in safety-related systems

Care should be taken in the selection of ML techniques, particularly when adopting pre-existing algorithms:

**(a)** The underlying statistical assumptions in the ML algorithms and their appropriateness for the particular problem data need to be understood and justified.

**(b)** Use of supervised learning techniques is susceptible to human error; conversely use of unsupervised techniques may not result in algorithms that are intuitive to human understanding, leading to difficulty in assuring the safety of the system.

**(c)** Inadequate data quality can negatively impact the accuracy of a machine learnt model; this includes the factors mentioned above, noisy and otherwise corrupted data (including the effect of hardware vulnerabilities and adversarial attack).

**(d)** Dynamic learning (including through learning from operational data) should be managed in a way that is consistent with current functional safety guidelines which require that, prior to change, an impact assessment is carried out and that, prior to re-introduction to normal operational use, the system safety argument is re-validated to take account of any change.

# Section 6

## Verification, validation and assurance

### 6.1 Background to verification and validation in safety-related applications

Verification and validation (V&V) are processes normally applied to any development project, regardless of whether the system, element or component is safety-related or not. In essence, verification is the process of developing confidence that the product meets its requirements and validation is the process of developing confidence that we built a product that meets the users' needs[1]. Sometimes, this is presented as two high-level questions: "have we built it right?" (verification); and "have we built the right thing?" (validation).

The processes used to undertake V&V range from analytical techniques to inspection to demonstration (through simulation and test), to confirm that the appropriate activities have been carried out competently (for example, by persons independent of the AI designers) and with sufficient rigour, and any issues identified have their root causes understood with appropriate measures taken to ensure that they are resolved. Conventional functional safety standards require that the V&V activities are carried out across the entire development lifecycle, with the initial lifecycle phases relying more strongly on analytical and inspection techniques and the validation phases more strongly on test and demonstration (repeated iteration of the phases is also possible, as, for example, in the case of 'agile' development). Due to the complexity of any real-world computer system, reliance on testing alone is not considered sufficient, and instead there are strict requirements for the use of various forms of analytical assurance during the design, to demonstrate that each stage of the design is without error or misunderstanding compared to the pre-curser level of design (for example, mathematical and/or logical methods, traceability analysis). Further, architectural strategies have gradually been accepted into common practice as a means to reduce the design complexity (for example, the architecture of the industrial programmable controller and the aircraft Full Authority Digital Engine Control (FADEC)). A third strategy is to use redundancy and diversity, including 'n out of n+1' voting, such that should an error be encountered in one branch of a system, then either through other branch(es) and/or the outcome of conditional voting, the required operation will continue to be provided. In some safety-related systems, temporal monitoring may be used to ensure that branches are unlikely to encounter the same error at the same time.

### 6.2 Verification and validation challenges

The challenges for safety-related systems developed using AI include:

**(a)** The phases of the AI lifecycle can be very different to those in current functional safety standards and as such, some of the existing analytical techniques (for example, logic) may be of limited value. Others (such as hazard analysis and risk assessment) may be alien to the way the AI discipline has evolved. Furthermore, whilst a significant level of consensus has grown up in the field of conventional safety-related system engineering as to the different methods and techniques that can prevent and control errors, given the relatively short time in which AI systems have been introduced to real world use, no such consensus is yet available to deal with the different phases of the AI lifecycle.

1 In IEC 61508, verification is defined as "confirmation by examination and provision of objective evidence that the requirements have been fulfilled." Validation is defined as "confirmation by examination and provision of objective evidence that the particular requirements for a specific intended use are fulfilled."

**(b)** The model of AI behaviour within a system is usually derived by statistical analysis of an observed set of individual data points. As such, the rationale for asserting that the model is correct and complete may be difficult for a human to understand or to relate to other well-proven views of the world; instead, heavy reliance is placed on 'demonstration' through the ability of the system to respond as expected to a defined test set. In practice, this addresses only a relatively small subset of the requirements of current functional safety standards, and, as already noted, is unlikely to enable the safety-related system to be demonstrated through test alone to the levels of integrity required by functional safety standards (probability of failure on demand between $10^{-1}$ and $10^{-5}$ or average frequency of dangerous failure between $10^{-5}$ and $10^{-9}$ probability of failure per hour, depending on the safety integrity level of each safety function).

**(c)** The paradigm adopted in AI is generally one of progressive improvement through observing and adapting to input data that varies from that seen previously. This is the antithesis of conventional safety-related systems, where extensive effort is put into assuring that the system will operate as required to maintain its safety function(s) from the outset; constant change is neither desirable nor practicable, since each change must be fully re-verified and re-validated, including consideration of any side effects of interfacing systems, operations or the environment.

**(d)** Demonstrating that AI cannot interfere, as necessary, with systems providing the primary protection and/or cannot exceed safe operating conditions.

Unfortunately, at the current time, there is no agreement on good practice to overcome the above challenges. The approach taken in [Ref 15] is to:

**(a)** If possible, apply the existing functional safety standards, at least to those parts of the safety-related system which are amenable to such processes as [Ref 15] has shown that even with an AI system, much of the system draws on conventional technology.

**(b)** (If **(a)** is not possible), identify equivalent safety principles to those underlying conventional functional safety standards and justify the equivalence in the safety argument.

An example of an alternative approach to justifying (parts) of the safety argument is given in [Ref 36] for the specific case of low-level verification requirements. [Ref 36] builds on [Ref 11] to include additional data specific elements as used in ML, i.e., that data should:

- **(i)** Relate to the intent of the high-level data passed to the ML from the system specifications.
- **(ii)** Not contain unintended bias.
- **(iii)** Be sufficient to achieve adequate model training.
- **(iv)** Be syntactically and semantically correct so that errors are not caused during training and misleading outputs are avoided during use.
- **(v)** Address normal and robustness behaviours. That is not solely focussed on nominal data that might only test conditions the AI is exposed to through routine operation, but also includes stressing but realistic inputs that may cause model failure.
- **(vi)** Be self-consistent, for example, the same type of item should not have different labels.
- **(vii)** Conform to standards for data used in safety critical applications (for example, [Ref 36]), which will assist with the correctness of the overall product as well as supporting re-use with other applications.
- **(viii)** Be compatible with the target computer.

**(c)** (If **(b)** is not possible), restrict the use of the AI to areas that do not impact safety. This may mean that approaches such as AI safeguarding (see Section 3.4) should be applied, or else strategies deployed whereby human oversight can be provided to avoid an unsafe condition arising from AI, for example, in the form of decision support tools or offline supervision.

Another example of guidance on assurance of autonomous systems can be found in [Ref 37].

Nevertheless, from the discussion on machine learning, it is clear that AI enables detection of anomalous behaviour (such as external attack) and automation of novel functionality, even control of phenomena not previously understood (for example, understanding the internal behaviour of novel combustion chambers which cannot be adequately instrumented without affecting the behaviour of the system itself). Hence, AI can certainly provide additional capability over that of conventional systems, potentially reducing 'out of bounds' events, enabling 'new' types of safety function and protecting against unexpected behaviour. In principle, AI does have the potential to increase overall safety, although this argument will need to be made at a higher level than that of the AI system itself.

## 6.3 Summary and conclusions

This section has introduced the concepts of verification and validation and has explained that for computer-based systems, test alone is not considered sufficient for safety-related applications. Instead, verification is applied throughout the design lifecycle using techniques ranging from the analysis to inspection to demonstration (through simulation and test) and to audit. A number of challenges for verification and validation of AI are noted.

## 6.4 Considerations for verification, validation and assurance of safety-related systems

When developing and implementing approaches to verification, validation and assurance for safety-related systems that are intended to incorporate AI, it is necessary to give attention to a range of factors that include:

- **(a)** Wherever possible, identify parts of the AI system that can be treated under conventional functional safety standards.
- **(b)** The approach to the overall assurance argument will need to be considered as it is unlikely that test alone will be sufficient to meet the requirements of current functional safety standards.
- **(c)** What analytical techniques, in addition to demonstration against a test set, can be deployed to improve confidence in the integrity in line with the expectations of functional safety standards.
- **(d)** In addition to confirming that the safety-related system satisfies its specification and the user requirements, it is also important to check that the safety-related system has not been compromised, for example, by performing some checks with altered or unexpected inputs.
- **(e)** In lower risk applications, if machine learning is allowed to take place after deployment (for example, through use of a separate monitoring system not able to influence the operation), re-verification and re-validation should be undertaken before each change is introduced into the operational system. In order to minimise rework, it may be possible to consider the use of a separate 'safeguarding' system to ensure that the AI cannot exceed safe boundaries of behaviour. Also, it may be possible to move towards a more permissive approach to re-validation when there is more evidence available from real applications.
- **(f)** Pay particular attention to the ways in which the data can result in incorrect learnt behaviour.
- **(g)** Ensure that AI cannot interfere with systems providing the primary protection, which depends on the results of analysis of the boundaries of the systems and any overlaps that may occur.

## Section 6 – Verification, validation and assurance

**(h)** Focus should be given to providing a clear understanding of the phases of the AI lifecycle and the differences to those in current functional safety standards, and evaluating whether or not some of the existing analytical techniques (for example, logic) may be of limited value whilst others (such as hazard analysis and risk assessment) may need to be adapted to reflect the way that the AI discipline has evolved. This approach is intended to determine if a consensus can be reached on measures and techniques to give confidence in AI robustness throughout its lifecycle.

# Section 7

## Security

### 7.1 Information security and the confidentiality, integrity and availability triad

The evolution of AI in safety-related systems has, to date, largely focussed on the potential benefits in identifying and defending against computer-based vulnerabilities and cybersecurity threats. But this should also prompt organisations to consider the security impacts of AI, including how AI technologies themselves can be secured, for example, ensure the integrity of decision making where this informs functional safety and the performance of safety-related systems.

In August 2022, the National Cyber Security Centre (NCSC) published *Principles for the security of machine learning* [Ref 38]. The principles of [Ref 38] should be considered alongside cyber security best practice for conventional software development.

A common model for considering the security of systems is the information security triad (CIA triad). Three principles (Confidentiality, Integrity, Availability) should be upheld if a safety-related system is to be secure. Depending on the application, different elements of the triad may be optimised. In networking, for example, cybersecurity would generally prioritise availability. In IT, confidentiality may be of paramount importance. In safety applications, integrity and availability (to perform the safety functions), are most important for system operation, though confidentiality is important to ensure no unauthorised access to the system and to ensure vulnerabilities (especially those affecting safety) are not widely known.

It is important to consider the different phases of the safety-related system lifecycle, any known vulnerabilities with the technologies selected for use within the system and their risks within the CIA triad.

### 7.2 Confidentiality

**Design:** All details of the safety-related system design and its data sets should be protected (including any AI frameworks used, the training methods, data libraries, inputs and model parameters). These can influence the implementation and ultimately the characteristics of that system. Additionally, any rules or conventions that the safety-related system follows, should be protected. This information could assist users with malicious intent, allowing them to degrade the operation of that system – for example, by providing input data that is known to trigger an incorrect behaviour or classification from the AI model.

**Training:** In training, data on the performance of the safety-related system, plus details of the actual data sets used (and any categories), needs to be protected during, and for, the lifetime of that system.

**Deployment/operation and maintenance:** All persons working on the safety-related system (during operation and maintenance) must understand the criticality of system confidentiality – details of that system, its specification and its performance, should be strictly confidential (if necessary, these could be bound by non-disclosure agreements). Only authenticated and authorised users should be able to access the safety-related system (both hands-on and remotely). Additionally, there should also be a limit to how much information can be gained from unauthorised access. Disclosure of the safety-related system's expected behaviour, limitations or the constraints by which it operates, could lead to persons taking advantage of how that system operates in order to trigger unexpected performance.

In practice, it is likely to be necessary to apply higher levels of authentication and authorisation to personnel with responsibility for safety-related system modification(s), as their role in undertaking these activities is key to maintaining system confidentiality. This extends to confidentiality of test data used within either the development or deployment environments to establish that performance is not undermined because of any changes to software or data introduced during modification.

**Retirement or decommissioning:** A safety-related system that utilises AI has almost certainly processed data and is likely to contain artefacts of that data or the processing results. If sensor data is stored, this may contain sensitive data (either personally identifiable information, intellectual property or deployment environment data) that is likely to be confidential and require removal from the safety-related system in order to protect it from falling into the wrong hands. Aside from the data, the safety-related system may contain implementation details of the model, or it may be possible to retrieve these, and the hardware of the system may need to be 'cleaned' to remove this.

## 7.3 Integrity

**Design:** Issues encountered during design include the re-use of components, their legitimacy and their integrity. If third party components (such as models, labelled data, software modules) are used, it is important to check that they are suitable (given the intended use), genuine and have not been tampered with. Failure to ensure the integrity of the components selected for use within a safety-related system can lead to inadvertent operation of compromised components that will not perform as expected or may have hidden functionality that could be triggered by particular input. These could all potentially result in unsafe and unpredictable behaviour.

Where third party components are used, it may be necessary to introduce measures to secure the supply chain using techniques to add resilience to AI models, organisational processes and personnel competence. Further information on this aspect of the security pillar for AI is provided in guidance produced by the European Telecommunications Standards Institute (ETSI) in their guides on securing artificial intelligence (SAI) series [Ref 39].

Similarly, arrangements should be made to improve the integrity of the infrastructure associated with the development environment and digital assets for AI in safety-related systems. This is necessary so that there is a clear understanding of the severity of the risk posed by the combination of valuable assets and supply chain inputs merging at this stage. Further information on measures that can be used are provided in [Ref 38].

**Training:** Likewise, the training data will need to be checked for its suitability, lack of bias and evidence of any tampering – this includes labels if present. If training data has been tampered with, it may contain malicious 'backdoors' which could allow users to trigger particular responses for inputs with particular characteristics. For example, sunglasses inserted into data used to train a facial recognition model. This altered training data can result in a model that associates sunglasses with a feature that it recognises, leading to, for example, an unauthorised user gaining access to a hazardous facility – endangering themselves and others.

**Operation and maintenance:** Systems that use AI rely on key 'ground truth' data (namely date, time and position information). If the source and accuracy of this is not protected (for example, if it can be changed without authorisation), then the correct functioning of a safety-related system cannot be guaranteed. Critical system operations such as the ordering of data, decisions made and detection of tampering, may perform erroneously with safety consequences. Key data should be protected at the appropriate privilege level. Ideally the safety-related system should have a fail-safe action if it encounters anomalous or conflicting data. Additionally, if operational data is able to modify the system's learnt model, consideration needs to be given to the risk of deliberate or accidental drift in the model. On the one hand, this could lead to improved performance, on the other, it could lead to a safety-related system that does not behave in accordance with its original objective.

**Retirement or decommissioning:** Continuing to protect the safety-related system once it has reached the end of its life will ensure that copies, or re-used/repurposed systems, do not circulate and impact the integrity of the in-service system or similar systems – for example, if 'version 1' or its constituent parts appear on eBay, this will impact the security of all other 'version 1' systems and potentially newer versions that share aspects of its specification.

This means that any safety-related system withdrawn from service needs protection to ensure that similar systems (now or in the future) are not put at risk by leakages of confidential design or operating features. Additionally, any data sets used in the system creation should be protected.

## 7.4 Availability

**Design:** Knowledge of the safety-related system and all of its components (versions, source, authenticity) is critical to ensure that the system does not contain any known vulnerabilities and to apply patches if vulnerabilities are disclosed in any of its components. The list of a system's software components is known as the software bill of material (SBOM). Such a list is important to ensure that it is easy to check for security vulnerabilities (as well as configuration management). Indeed, the US Government has now mandated [Ref 40] that providers of critical government software supply an SBOM for each product and it seems likely that other countries will follow this lead.

At the design phase, mechanisms should be included to verify all inputs to the safety-related system – ensuring that they are in the correct format and within range. If the safety-related system receives external data, then its reliance on this data should be assessed to ensure that risks are known. If external data is critical to the correct operation of that system, then checks on the authenticity of the data will be needed.

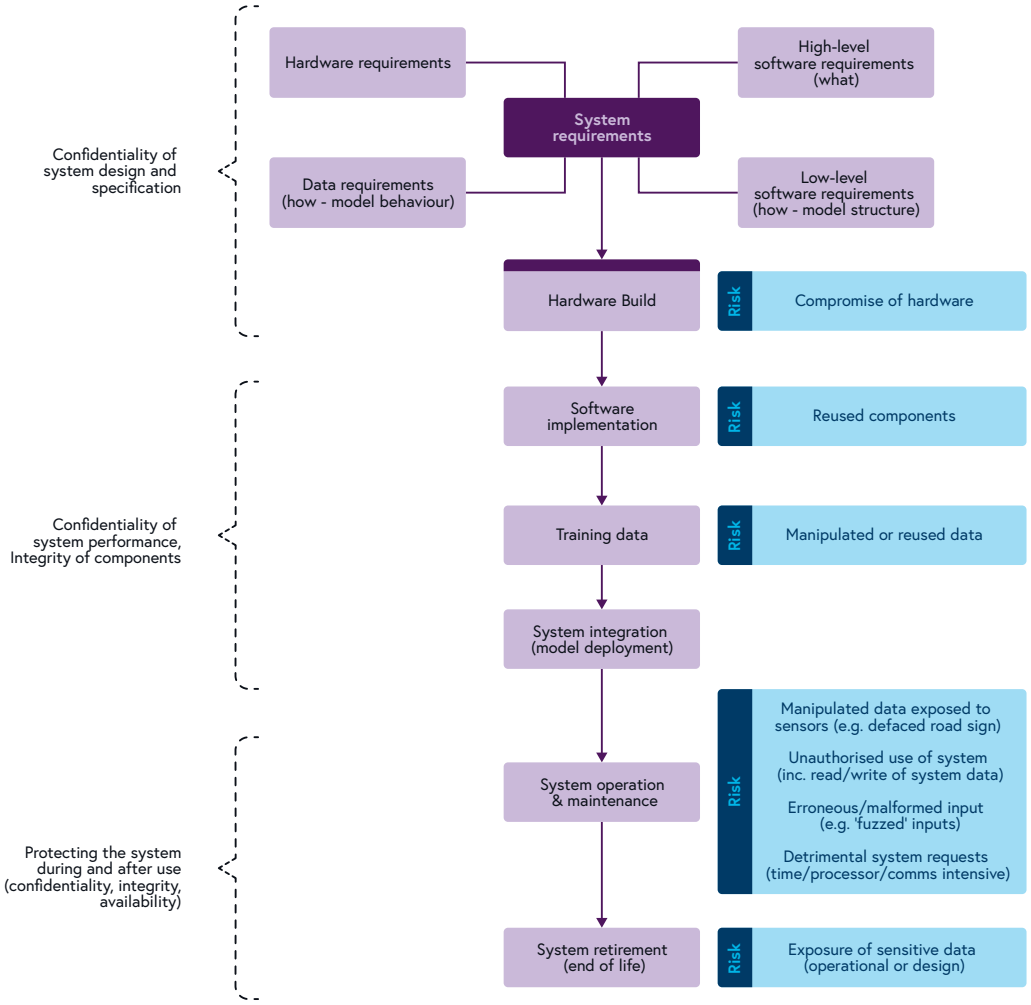
**Operation and maintenance:** Access to a safety-related system needs to be restricted to only those that have a legitimate need to use it for their role. Users should be suitably knowledgeable and skilled to use the system correctly and determine inappropriate use (for example, in environments that the system has never encountered). They should also be alert to any indications of malfunction or unexpected behaviour – especially if safety is impacted. Additionally, the safety-related system should be protected from denial-of-service attacks and/or computationally 'greedy' operations, demanding frequent, time consuming or resource heavy tasks. In addition, an access control system should be in place such that each user only has permissions to carry out the activities they need for their role. Failure to prevent unauthorised access can impact all three elements of the information security triad and if actions performed are detrimental to the system's functioning, then it is likely that safety will also be put at risk.

**Retirement or decommissioning:** Continuing to protect the safety-related system as it is decommissioned will ensure that existing models of that system that are still in use are not put at risk by knowledge of the system's operation, or by re-used or non-genuine models.

## 7.5 Sources of security threats throughout the AI lifecycle

A system that uses AI involves additional considerations above and beyond those that would be expected for the secure development of a traditional software-based system. These are particularly important for a safety-related system. The diagram shown in Figure 7.1 highlights some of the security threats and areas that require protection within a system using AI.

**Figure 7.1** System lifecycle for a system incorporating AI: Potential risks are shown with purple arrows. Braces indicate elements of the Computer Security CIA triad




## 7.6 Competence and access control

In addition to security concerns, competence should be addressed throughout all lifecycle phases. The knowledge and experience of those operating the safety-related system as well as those that design, develop, train and deploy systems using AI in safety-related applications, will influence the decisions made across the system's lifecycle, and the level of risk that is deemed acceptable for the intended application. Knowledge will be required in the domains of security, AI technology, functional safety and the intended operational environment. The level of competence required for each role needs to be assessed and individuals verified for those competencies.

Competence should be taken into account in authenticating and authorising personnel to operate and use, maintain and, as necessary, modify safety-related systems that include AI. It is recommended that only authenticated and authorised users should be able to access and use the safety-related system, and only a subset of personnel will be authorised to modify system settings. Verifying competence is often dealt with through a competency management scheme, which sets out the competencies expected and evidence required to prove competence in specific tasks, including schemes for monitoring and measuring the competencies of employees [Ref 41].

Depending on the application, a user could need competence in the principles underlying the assurance case for the particular system, as well as the AI assumptions underlying its model. A user that can correctly interpret whether a system is performing as expected, or malfunctioning, will be better positioned to judge the system's output. Alternatively, the user may only need to know the constraints on use, in which case the information should be provided in the form of a 'safety manual', data sheet or operations and maintenance manual.

In terms of competence requirements, preventing the system encountering anomalous input (or detecting it and responding with appropriate behaviour) depends on an understanding of the principles and constraints by which the system operates, including any situations where it might not work as intended.

Autonomous systems enabled through AI may not have an operator and therefore may have need for protection beyond those of conventional safety-related systems. The first dimension to this is physically protecting the safety-related system when it is not in use – preventing tampering or theft, logically preventing external access and authenticating any changes. The second element is controlling input into the safety-related system's sensors – this could be achieved via filter, or other guard mechanisms. For example, preventing intense light beams from penetrating camera sensors.

Finally, care needs to be taken to ensure that documentation relating to all aspects of data management are subject to appropriate processes, such as organisational controls relating to access and modification, to prevent the introduction of inadvertent data errors that may not be readily detected until later phases in the safety-related system's lifecycle model. This may be achieved in practice through the adoption of relevant systems engineering principles in the lifecycle model and, where necessary, adapting these to meet objectives for security informed safety.

## 7.7 Summary and conclusions

Security and safety are interrelated – a system cannot be guaranteed to be safe if it is not secure [Ref 42] – although it should also be understood that a secure system is, by no means, necessarily safe. This section has provided a summary of how security needs to be considered at all phases of the overall safety lifecycle, regardless of the technologies used by the safety-related system and its development and deployment environments. This necessity has been termed "security informed safety" [Ref 6].

## 7.8 Considerations for good security in safety-related systems

Only by considering security during all phases of the system lifecycle, can the security of the safety-related system be protected through life - this is a basic pre-requisite for demonstrating that a system is safe:

- **(a)** Follow cyber security best practice – the underlying system is a computing system – regardless of its use of AI.
- **(b)** Design for security.
- **(c)** Minimise the availability of data on the safety-related system (to minimise an adversary's knowledge).
- **(d)** Secure the supply chain and both the development and deployment infrastructure.
- **(e)** Track your digital assets.

# Section 8

## Algorithmic behaviour

### 8.1 Differences between conventional and AI algorithms

This section outlines the challenges associated with algorithms used to produce AI like behaviour. In the context of this publication, an algorithm is defined as a process or series of steps that can provide a solution to a problem. In AI applications, particularly those that make use of ML, the steps of the algorithm are usually known but the specific behaviour at each step is dictated by additional inputs. Examples include the application of training data and the tuning of parameters via the learning algorithm (in the case of deep neural networks) or the use of pseudo-random behaviour (for example, during the explorative behaviour of numerical optimisers such as genetic algorithms). In general, an algorithm provides a means to map an input domain to an output domain, but in the case of AI, the mapping is usually so complex that it is difficult to make claims about the predictability in the same way as has been argued for conventional safety-related systems.

In the case of conventional functional safety standards, assurance is built up from the initial derivation of the safety requirement(s), through traceable decomposition through each successive behavioural specification and onto implementation. At each phase of the development lifecycle, different activities are undertaken, each seeking to prevent, detect and remove errors as close to the source of their potential introduction as possible. Due to complexity of the resulting algorithms, it is recognised that verification cannot be exhaustive and for all but relatively simple algorithms, it is impossible to demonstrate the absence of errors. Fortunately, this uncertainty is mitigated by the predictability of the algorithmic behaviour across a known input data space.

For an AI derived algorithm there are two main challenges:

1. Demonstrating that the algorithm does predictably implement the expected behaviour.
2. Demonstrating the ability to control any hazardous situation that may arise as a result of using the algorithm.

In AI algorithms, predictability is often lost for a number of reasons. For example, when using ML to derive the algorithm, the resulting behaviour is dictated by the data it encounters during any learning phase. This contrasts with traditional systems that are explicitly derived from a higher-level specification and directly encoded. In some ways, ML is analogous to the traditional scientific approach where a model of the world is derived through observation, albeit in a much less-well defined sense. Further, an algorithm derived by ML will include the ambiguities of the learnt environment. Additionally, AI-based systems are specifically designed to cope with dynamic, complex and uncertain environments. Hence, it is not practicable to define the exact behaviour for every instance of the operational environment (indeed, if that is so then arguably the software should be encoded using traditional, predictable methods).

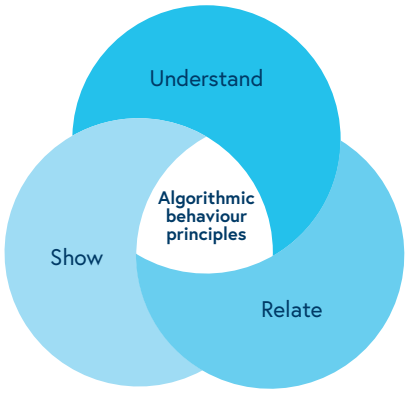
In conventional automation, a system will be developed to provide a benefit to the intended users, for example, to travel from A to B in a specified time or to produce energy at a specified level of consumption elsewhere. In general, the benefit may not be directly related to safety, but in many cases, the use of the system could increase the safety risk. Accordingly, it is necessary to limit the risk to that which is tolerable [Ref 43] and As Low as Reasonably Practicable (ALARP) [Ref 44] in consideration of the overall societal benefit. This may be achieved, either through the inherent safety of the primary system or through provision of supplementary 'protection and mitigation' algorithms. In functional safety, these requirements are derived from, and therefore directly traceable to, the ways in which the algorithm can lead to an unsafe situation. In contrast, an algorithm derived through ML has no such linkage to a requirement based on control of a dangerous state, since such a state could never be allowed in normal use. Instead, behaviour would need to be derived indirectly from identification of constraints that cannot be allowed to be breached.

## 8.2 Assurance of algorithmic behaviour

Assurance of algorithmic behaviour can be distilled into three basic principles:

1. **Understand** what the system needs from the algorithm. In a traditional system, this might be considered as the high-level system requirement; that is "what the system requires the AI to do", rather than how it is to achieve it.
2. **Relate** the understanding of system-level requirements to the implementation of the algorithm. This can be considered as broadly equivalent to low-level requirements in traditional software development (that is "how the AI will achieve the requirements the system is placing on it"). In traditional safety-related software development approaches, this is achieved through traceable decomposition. This may not be practicable in many AI-based implementations, in particular those involving ML.
3. **Show** the implementation is appropriate. This can be a combination of:
   **(i)** arguing that the process used to implement the behaviour is credible (for example, making claims that assurance of algorithmic behaviour can be distilled into three basic principles, and that the AI-paradigm is the most appropriate for the type of problem being solved); and
   **(ii)** demonstration through the application of verification and validation that the probability for errors, when applied in the target domain, has been minimised.

**Figure 8.1** Model showing the relationship between the three basic principles of assurance of algorithmic behaviour




Whilst the three basic principles may appear to be sequential, they are most likely to form an iterative lifecycle.

## 8.3 Summary and conclusions

This section has summarised the challenges associated with algorithms used to produce AI like behaviour and, in particular, the lack of explicit traceability between an algorithm's behaviour and that of both the user need and the control of potential hazards. Three principles of assurance of algorithmic behaviour have been proposed:

1. understand what is required from the algorithmic behaviour;
2. relate the understanding to the control of hazards (in the case of safety) and, more generally, to the user need; and
3. show that the behaviour executed is desired, and that the safety-related system cannot execute any behaviour that might jeopardise safety.

## 8.4 Considerations for AI algorithmic behaviour in safety-related systems

Care should be taken in the choice of algorithm to deliver both operational and non-operational (for example, safety) requirements including:

**(a)** Ensure the selected algorithm is appropriate for the range of tasks expected of it.
**(b)** At the (global) level, ensure the general behaviour of the algorithm can be explained and related both to real world models of the expected behaviour and to the expectations of safe behaviour. Ensure that any constraints on behaviour for safety reasons cannot be breached.
**(c)** At the finer granularity (local) level, ensure there is a thorough understanding of what a particular output should be and why it occurred.
**(d)** Consider the extent to which the algorithm choice should support the identification and analysis of incidents, weaknesses or unexpected conditions. In the event of such a condition, ensure that the algorithm will bring the safety-related system into a safe state.
**(e)** Ensure protection has been provided for typical algorithmic errors including over-fitting, data leakage, under-representation and adversarial examples.
**(f)** Ensure the programming language supports the data used within the problem space.
**(g)** Ensure that an argument for the dependability of the supporting software libraries and their dependencies is available, including that they come from reputable sources and that there is good knowledge of any unexpected or failed behaviour. Ensure that the specific configurations are recorded and that they are likely to be supported throughout the life of the safety-related system.
**(h)** Ensure that the development environment and support tools used during the software development have sufficient pedigree, are properly documented and can support any particular AI considerations. Wherever possible, use conventional functional safety standards to ensure tools are qualified to adequate safety integrity levels.
**(i)** Software failure – algorithmic issues can be less visible than traditional software faults; ensure measures been implemented to provide confidence in the AI model under failure and abnormal operating conditions.

# Section 9

## Human factors in AI safety

### 9.1 Overview of human factors (HF)

Human factors (HF) describe the complex interactions between humans and built systems, whether they be physical interactions, mental models or supervision and control. HF considerations include the recognition that it is possible to have designed and built a safety-related system which satisfies the specified safety requirements, but which, when deployed, exhibits undesirable behaviour owing to the way actors interact with that system. As with functional safety standards, the attributes associated with strong design for human factors in traditional safety-related systems are relatively well tried and tested over the course of many decades.

### 9.2 Impacts of AI on human factors

The implementation of AI in safety-related applications is likely to require a re-evaluation of the tried and tested HF processes. As the capability of AI, and hence, autonomous behaviour, increases, the boundary of behaviour may shift between the human and the machine, changing the traditional activities that the human may be expected to carry out. It might be assumed that a commensurate reduction in supervision would yield fewer HF related issues, however, it is more likely that the point at which HF considerations need to be addressed will change. For example, risks may shift from the operational phases into development as unrealised HF risks become 'baked in' the AI model, rather than manifesting as a result of in-service operation. Other considerations that may need to be taken into account include:

**(a)** the extent to which a human operator will place dependence/trust on an autonomous machine; and
**(b)** the extent to which, in the case of unexpected behaviour, the human is in a position to understand the situation and to take control to maintain a safe state.

Further, if the operation of a safety-related system is largely autonomous, then a human operator is likely to be subject to little practical experience during various modes of operation; training for abnormal and exceptional conditions will be necessary if it is not possible automatically to bring that system to a safe state.

Some examples of potential HF risks are discussed below:

**(a)** the limitations of effective decision-making capability in AI models when compared to humans;
**(b)** human factors risks associated with incompatible understanding of system operations;
**(c)** issues related to differences in human culture compared with the context in which an AI model is used; and
**(d)** management of the human/machine capability gap and its effects on those developing safety-related systems.

## 9.3 Performance goals for AI models based on the same criteria as for human-centric systems

In certain applications with a limited requirement for safety integrity to be achieved, a task may be defined by implicit criteria and left to the operator to 'fill in the gaps', for example, "the task is complete when the object is placed in this area." Despite the incomplete nature of the goal, human cognition, experience and problem-solving capability mean that the goal is readily understood by operators and developers alike due to a common contextual model of the world. However, when combined with certain ML techniques such as reinforcement learning (RL), the lack of a complete specification is likely to lead to behaviour not anticipated by the system designer. In particular, despite terms such as neural network, an ML algorithm only has access to the data presented during training from which a deterministic algorithm is derived. It does not have access to a wealth of general knowledge of the type used in human problem solving. In short, expecting humanlike approaches from ML systems when exposed to ill-defined or unexpected conditions is to be exposed to risk.

Such shortcomings in background experience will require increased rigour from system designers in addressing the gap between desired system behaviour and the type of behaviour they would expect an equivalent human to exhibit. This is not a simple change to make. It requires a change of mindset from human to machine-centric design, which in turn needs to be codified into design practices to ensure they are not overlooked.

There are approaches, such as [Ref 26], which address the definition and validation of system behaviour. Peer review can also be a useful technique to identify ambiguous and alternative interpretations of a specified objective.

## 9.4 Humans acting on incomplete information

In both conventional and AI systems, fall-back to human management is used as a control measure in the event a safety-related system approaches an unsafe state. Humans have a relatively good track record with providing timely and appropriate responses to dangerous situations in a way that seeks to reduce harm as effectively as possible.

Traditional safety-related systems make use of tried, tested and often safety assured components to either intervene or warn supervisors of unsafe states, through mechanisms such as aural/visual warnings or critical parameters displayed on human machine interfaces (HMIs).

Since scenarios where humans are required to intervene are likely to present high stress, high cognitive load and conditions that are rarely experienced; careful thought goes into prescribing the flow of events leading up to the dangerous state and the appropriate actions to be taken. This includes the time for the human to acquire the necessary situational awareness. History shows that many accidents are attributed to human error, because the human was not able to adapt to the unexpected operational conditions sufficiently quickly.

Systems using AI lend themselves well to increased autonomy which naturally requires less by way of HMI design, indicators, etc. Further, it may be difficult to predict the onset of hazardous states and to articulate the situation to supervisors effectively, even if the appropriate alarms etc. are present.

Designers must consider the likely mental models that operators and supervisors have of any system, both in normal and failed states. Warnings should be generated such that they require minimal understanding of the way in which an AI system processes and generates data. Designers should also seek to create defined bounds by which control can be handed over to the operator, ensuring that for any foreseen failure state, there is a human response that can place the system into a safe state in sufficient time to avoid an accident.

The use of specialists in HF, particularly those in branches such as cognitive ergonomics, may be of importance in the early-stage design for developing safety-related system states, ensuring humans in the loop are able to act in the manner desired by the system designers.

## 9.5 Culture and context

The practice of functional safety is largely harmonised throughout the industrial world due to the wide acceptance of standards such as [Refs 6, 7 and 11]. However, the concept of tolerable safety risk can vary greatly between different communities and different sectors. Different cultures, with different accepted 'norms', behaviours and risk appetites, may act very differently when faced with the same situation. Another variant is the level of awareness/capability that can be expected from the intended users. For example, while operators within the high hazard industries (for example, nuclear, aviation, energy production) can be expected to be highly trained, alert and responsive, it is important to acknowledge that not all users will have the same level of experience and proficiency. Additionally, users may not be safe at all times – for example, if they are distracted due to human factors, issues such as temperature, noise, fatigue or illness. Furthermore, there are traditional industry sectors, especially workers in small to medium enterprises, who are not familiar with a functional safety approach, but may have adopted AI with safety implications.

AI enables a whole new raft of applications, for example, in medical devices, mobility, personal assistants and other forms of assisted living, which in turn will introduce new sources of hazards (for example, slips, trips, falls). It is important that such safety-related systems are appropriate for their intended users and that any vulnerabilities (for example, externally connected home devices, inability to control mobility devices) be properly controlled. Furthermore, with the potential for autonomous behaviour to trespass into the realms of ethical decision-making, it is important to recognises that different cultural and legal jurisdictions may require different responses to the same situation.

## 9.6 Integration and utilisation of new engineering disciplines

As the use of AI becomes more prevalent, developers and operators of systems may need to draw upon new skill sets and talent pools to effectively develop their systems. This is particularly reflected in the increased emphasis placed on the importance of disciplines such as data science and ML programming, among others.

Historically, mechanical, electrical and control systems engineers have been integral to the design of safety-related systems, having a strong basis of historical examples on which to develop new safety-related systems, and many deployed safety-related systems to learn from. This is often based on a long history of learning from accidents, for example the rail and aerospace industries amongst others. Data scientists, programmers and other AI-centric disciplines need to ensure that they appreciate the safety concerns that are fundamental to the use of their systems in the real world. In particular, they need to appreciate the impact that an incorrectly tuned algorithm can have on the output of a system that could either directly affect safety, or alternatively, put increased demand on a safety-related system that is responsible for maintaining a safe state.

Further, to ensure the appropriate use of AI systems, the lines of responsibility between human operators/users and the actions of the AI need to be understood and clearly articulated. There are two aspects to this – firstly, the need for a responsible person to oversee that the safety-related system is only used for environments and contexts in which it is known to operate correctly. Secondly, the need for an operator to understand what situations they need to be alert to in order to rectify anomalous behaviour and ensure that it does not affect safety. The security pillar touched on the competence of users and operators working with safety-related systems that utilise AI. From a safety perspective, existing regulations such as COMAH [Ref 45] for systems where dangerous substances are in use, ROGS [Ref 46] for guided transport systems such as railways and tramways, and CDM [Ref 47] for construction, design and management need to be considered, as will any incoming guidance for the use of AI and autonomous systems.

For AI, as with all new technology, organisations need to ensure that the new disciplines and skills of AI are recognised for their ability to improve system safety, yet scrutinised against historical errors to avoid making the same mistakes. This is particularly true when knowledge of past errors is codified deep within specific engineering disciplines and may no longer be explicitly recognised.

## 9.7 Summary and conclusions

This section has presented a non-exhaustive list of the HF challenges presented through using AI in safety applications, many of which are not yet completely known. Organisations can, however, address such uncertainty by ensuring that the impact of technological and organisational decisions on safety are properly considered during the design and development of safety-related systems. This can help to ensure that the downstream effects of AI on human factors can be controlled.

## 9.8 Considerations for human factors in safety-related systems

The role of human factors in relation to the impact of AI on both the safety-related system and the personnel involved in all aspects of its development is:

**(a)** Ensure that the effect of changing the boundary of responsibility for decision making between human and machine is properly considered during design.

**(b)** Consider whether the decision-making algorithms require implicit human management in cases where a machine is unable to acquire sufficient experience to act safely.

**(c)** Ensure that the mental model of the machine operation is comprehensible to, and understood by, the human supervisor; ensure that they are provided with adequate information in a timely manner to ensure a safe state is maintained.

**(d)** Consider whether additional training or support from either AI or functional safety expertise is necessary as the level of autonomy increases.

**(e)** Ensure the expectations placed on the user of the safety-related system are consistent with their experience (including technological awareness and cultural background) and capability (including consideration of their physical and mental alertness); ensure that proper consideration is given to ethical challenges associated with the design and use of the system.

**(f)** Be aware that the safety engineering and the AI disciplines have, to date, rarely experienced the other's field; it is important that both disciplines understand the strengths and limitations of AI with regard to safety and that the use cases to which it is applied are appropriate.

**(g)** AI technology has developed to such an extent that it is readily available; however, the basic competency requirements for safety-related applications [Ref 6] will apply, including an understanding of the AI technology limitations as well as the need to justify the safety argument.

# Section 10

## Maintenance and operation

### 10.1 The importance of maintenance

Maintenance is the continued monitoring and adjusting of the system operation to ensure that it continues to operate as expected, i.e., continues to satisfy its requirements specification, during the course of its operational service life. During a system's operational service life, it may need to deal with different sources of input data, as well as issues noted in earlier sections of this publication, such as data drift, changes in the environment, failure or errors in the wider system, and changing inputs arising from adversarial action. This section looks at considerations required to ensure that safety-related system integrity can be maintained during operating and maintenance activities.

Maintenance of a safety-related system can range from updating hardware components (such as sensors or actuators), software components (including bug fixes, operating system updates), or changes to the machine learning model itself. The latter may require re-training and re-testing to ensure that it continues to operate as expected given changing environmental, or other influences. Since the input to a system utilising AI is typically based on one or more sensors, these need to be maintained to ensure that they continue to function at the level required. For example, solutions exist in the automotive sector, which provide software automation of sensor maintenance tasks such as calibration, testing and replacement.

Regardless of the maintenance activity, any alteration to a system having safety implications will require an assessment of the impact of change, including side effects on other systems or operators. It will almost certainly need to be re-calibrated and re-configured to adjust for any changes; such re-calibration will need to be shown to be within safe bounds. It will then need to be tested, re-verified and re-validated to ensure that errors have not been introduced, and check that the safety-related system continues to operate as expected.

Maintenance can be a high-risk activity, both potentially compromising safety integrity due to error during modification or use of unauthorised software or hardware components, or through breach of security concerns. Careful control is needed to prevent disclosure of details of a safety-related system's physical or logical components, their configuration and their capabilities, since this could open a route to adversarial attack. There is also the potential for unintended mistakes, for example, by downloading a configuration not intended for the particular machine.

### 10.2 System integrity and confidentiality considerations

It is important that the wider impacts of operation and maintenance on the performance of a safety-related system that utilises AI are considered during relevant planning phases. These include:

- **(a)** checking any new components for legitimacy, integrity and functionality;
- **(b)** maintaining an awareness of supply chain risks; and
- **(c)** limiting the potential for deliberate or accidental introduction of unsuitable components.

Additionally, the confidentiality of a safety-related system needs to be maintained during its entire lifetime, including decommissioning. If details of that system and its capabilities are disclosed, then all operational systems of the same specification are likely to be at risk of adversaries utilising system knowledge to their advantage. This can range from deducing characteristics or parameters of the machine learning model, exploiting its weaknesses, physical limitations or methods of interaction. For example, a safety-related system's interoperability bounds originate from the rules of the domain in which that system is intended to operate, including the region, country and its environment. Examples within the UK include rules of the air and the UK Highway Code. These are the rules under which a safety-related system needs to operate, i.e., how it will behave when faced with particular circumstances in a given domain. If adversaries

have knowledge of the behaviour then they could exploit any weaknesses, such as where pedestrians may walk in front of an autonomous car - here the impact may be negligible, but if a system is misused due to its known reaction to a given set of events then there could be potentially adverse consequences for the system safety. Mitigations include careful selection of individuals responsible for aspects of system operation and maintenance, and training to explain the consequences of poor maintenance and operation. Where required, this could be backed-up with confidentiality agreements to limit the likelihood of information leakage.

## 10.3 Data considerations

During operation, a safety-related system will need to ensure that it is robust to inputs from all data domain spaces. It will also need to be able to detect and deal with data drift (see data pillar for more information). Additional re-training and testing of the AI model may be required in order to deal with changing environmental data, although this will also require a safety impact assessment and rework of the safety assurance case. In doing so, it will be important to consider the cause of the data drift, since it may be the result of an adversary deliberately feeding the model biased data in order to influence future decisions.

## 10.4 Users' critical role in maintenance - use of standard operating procedures

When considering the use of AI in safety-related systems, it could be argued that the AI element is influenced by that system's properties and its current state. Activities involved in maintaining the safety-related system itself can alter both of these, and there are certain activities that can be performed to maintain that system in a valid and functional state. To this end, the use of standard operating procedures (SoPs) can inform all users what activities they need to perform, how they are to be performed and when they need to perform them. By specifying activities required, such as:

- **(a)** regularly 'cleaning' the safety-related system (removing and archiving operating data, log files and user data);
- **(b)** ensuring that unknown media is not inserted into the safety-related system;
- **(c)** administrator-level activities, such as ensuring that user accounts are current, and password policies are suitable for the security of the operating domain;
- **(d)** regularly applying security patches and updating virus protection with the latest definitions (but consistent with the safety policies on re-assurance of the system after change); and
- **(e)** configuring the safety-related system according to the least privilege principle (i.e., the user is only given the privileges needed to perform their tasks).

These activities should be complemented by maintaining information on AI vulnerabilities and ensuring robustness to challenges that may arise throughout the overall lifecycle. This can be considered as an essential aspect of lifecycle management to achieve trustworthiness for AI components given that, in practice, demonstrating their integrity is likely to go beyond functional testing and conventional assurance methodologies.

Additionally, in order to ensure that a safety-related system is not used in situations where it has not been tested, or when it is known to operate erroneously, it is critical that users are aware of what that system can and cannot do, the domains in which it is designed to be operated, and those for which its operation may be unknown. Without an understanding of the correct use of the safety-related system and the interpretation of its outputs, users may put themselves or others at risk. The reader is referred to the human factors pillar where responsibility is discussed. Finally, during safety-related system decommissioning, care should be taken to remove all operational data from that system and any configuration, parameter, log file, or other information that could compromise similar safety-related systems still in use.

## 10.5 Summary and conclusions

This section has looked at the considerations for ensuring that an AI system's safety and security is maintained during operation and maintenance activities. The safety integrity of the safety-related system can be undermined if the operational domain changes, such that the validity of that system's training data is affected. It can also be undermined by faults that develop in the hardware of the system or by cyber-attack. Consideration must be given to how any changes (whether from external constraints or internal defects) will be managed and re-assured. Further, it is important to ensure that maintenance activities do not leave the safety-related system in an un-usable and/or a state that cannot be re-assured, since this would jeopardise the availability of a safety function, and therefore potentially of the operation.

## 10.6 Considerations for good practice during operation and maintenance in safety-related systems

The following factors require consideration for operation and maintenance of safety-related systems which incorporate AI elements, including:

**(a)** Maintain an awareness of supply chain risks and seek to limit the potential for deliberate or accidental introduction of unsuitable components.

**(b)** Check the legitimacy, integrity and functionality of all new or changed components.

**(c)** Prioritise the authorisation access to the safety-related system throughout the entire lifecycle – consider careful selection of individuals responsible for all aspects of system maintenance, and the use of legal agreements to limit the likelihood of information leakage.

**(d)** Mandate the use of standard operating procedures for all safety-related system users and activities, and ensure these are sufficiently robust and address security.

**(e)** Ensure that the safety-related system maintainers and users are aware of what that system can and cannot do, as well as the domains in which it is designed to be operated and those for which its operation may be unknown.

**(f)** Ensure that any change during operations or maintenance is subject to an impact assessment and properly re-assured (see Section 6 on verification, validation and assurance).

# Section 11

## Legal and ethical considerations

### 11.1 Legal implications of AI

AI can be considered as a key enabling technology for autonomous systems to operate in complex environments. However, as with many novel technologies, where it is used, it can bring legal and ethical challenges that may be less well understood and controlled than those associated with more conventional technology.

Due to its relatively recent deployment in real world applications, the legal aspects associated with the deployment of AI/ML are still emerging.

In 2021, the European Commission proposed to introduce a common regulatory and legal framework for AI [Ref 48] that targeted mitigating the risk of AI failures and the lack of trust by introducing a risk-based approach to regulating AI systems. This provides rules on the implementation of AI, including what are considered high risk systems. This includes systems where there is a risk of harm to health and safety. Initiatives are ongoing within the EU, standards bodies and several national and international bodies to produce standards that support this new legislation or address gaps in the adaptation to existing legal frameworks [Refs 48, 49, 50, 51 and 52 and 15, 16 and 17 to name a few].

Although the overarching legal frameworks are expected to be consistent with those applicable to conventional technology and are generally risk-based, there can be considerations that explore the bounds of responsibility between human and machine and also between development and use.

The UK is taking a different approach as set out in a white paper on a pro-innovation approach to AI regulation, which sets out its intention not to introduce AI specific legislation or a specific AI regulator. Instead, the intent is to establish a set of cross-sectoral principles to guide how regulators should approach common risks relating to AI, with regulators asked to interpret and apply these to their sectors, on a context specific basis, using existing legal frameworks.

Those creating AI systems need to understand the legal framework they are operating under, how it would apply to their system and the applicable legal requirements for compliance.

Key areas to consider include the duties of those in the supply chain under product safety legislation and the duties of those using the products under health and safety or other relevant legislation. AI supply chains can be complex and need careful consideration of accountability and liability.

The nature of AI brings some additional challenges in this respect, such as:

**(a)** Who is responsible for newly learnt behaviour after the point of supply?
**(b)** What is the providence of the data used in the system, and who is responsible for any errors?
**(c)** Who is responsible for the machine learning algorithm and its predictions?
**(d)** How will a supplier of a product provide the required information to the end user for them to maintain safety?

These questions are application specific and should be considered by developers and end users of AI systems which could impact safety.

Companies have duties under both civil and criminal law, and so do people. Individuals can have specific duties and, for example, under the Health and Safety at Work etc Act [Ref 53], employees need to take reasonable care for the health and safety of themselves and other persons.

Under civil law (tort), there is a right to recompense for those who have suffered loss or harm due to the wrongful or negligent acts of others (either individuals or companies). If an AI-based system is used with human oversight and the AI causes harm without the operator's intent or express permission, then should the operator be considered negligent? In such cases, the victims of the breach may have a less obvious path to legal action compared to a system under human control, even though the adverse impact may be the same.

One of the consequences of changing the boundaries of responsibility between the human and machine is that accountability needs to be identified and clearly assigned, during both design and operation.

Stakeholders of AI-based systems with autonomous behaviour and those providing safety-related advice need to be aware of the wider legal ramifications of the use of AI.

## 11.2 Ethical implications of AI

Ethical considerations come into play when AI is used to make decisions on the relative value of different courses of action. Examples include that of an autonomous vehicle facing a decision between alternative accident scenarios (for example, whether to change direction when faced with two equally likely potential crash options if a child runs unexpectedly into its path), or a catastrophic event requiring decisions to be made on where to bring down an aircraft. More subtle decision-making examples involve the design of the autonomous vehicles, for example, whether to optimise the safety of its passengers, the safety of third parties or the safety of the environment.

Given such challenges, a subfield of applied ethics, AI ethics, has emerged as a response to the range of individual and societal harms that the misuse, abuse, poor design or negative unintended consequences that AI systems may cause. AI ethics is developing a set of values, principles, and techniques that employ widely accepted emergent standards of right and wrong to guide the development and use of AI technologies. These values, principles and techniques are intended both to motivate ethical practices and to prescribe the basic duties and obligations necessary to produce ethical, fair and safe AI applications.

Ethical principles and guidance focus on the following themes:

**(a)** **Privacy:** Principles under this theme stand for the idea that AI systems should respect individuals' privacy, both in the use of data for the development of technological systems and by providing impacted people with agency over their data and the decisions made using the data.

**(b)** **Accountability:** This theme includes principles concerning the importance of mechanisms to ensure that accountability for the impacts of AI systems is appropriately distributed, and that adequate remedies are provided for the potential impacts.

**(c)** **Safety and security:** These principles express requirements that AI systems be safe, perform as intended (and do not exhibit unintended dangerous behaviour), and are also secure, resistant to being compromised by unauthorised parties.

**(d)** **Transparency and explainability:** Principles under this theme articulate requirements that AI systems be designed and implemented to allow for oversight, including through translation of their operations into understandable outputs and the provision of information about where, when and how they are being used in a way that still respects intellectual property.

**(e)** **Fairness and non-discrimination:** Acknowledging the concerns that AI bias is already impacting individuals globally, fairness and non-discrimination principles call for AI-based systems to be designed and used to maximise fairness and to promote inclusivity.

**(f)** **Human control of technology:** The principles under this theme require that important decisions remain subject to human oversight, noting that before deciding on a course of action, there needs to be sufficient understanding of potential impacts and sufficient time to prevent unintended consequences.

(g) **Individual and corporate responsibility:** These principles assert the role and responsibilities that individuals and organisations involved in the development and deployment of AI systems play in the systems' impacts, and call on their professionalism and integrity in ensuring that the appropriate stakeholders are consulted and the short- and long-term effects of decisions taken by AI (including potential negative impacts) are recognised and planned for.

(h) **Promotion of human values:** Finally, human values principles state that the ends to which AI is devoted, and the means by which it is implemented, should correspond with our core values and generally promote humanity's well-being.

There are a number of published frameworks associated with the legal and ethical adoption of AI-based systems [Ref 54]. The UK Government has published its own guidance framework [Ref 55], which itself leans on other frameworks such as the Data Ethics Framework [Ref 56].

In addition to these frameworks, in 2021 the EU Act [Ref 48] listed a number of prohibited AI practices to ensure the development of ethical AI systems that align to fundamental laws. These include practices such as the use of real-time biometric identification systems in publicly accessible spaces for law enforcement, and the use of AI for subliminal techniques to distort behaviour. High risk systems are required to be assessed by implementing ethics-based auditing using the conformity assessment procedure for AI systems [Ref 37].

## 11.3 Summary and conclusions

This section has looked at the legal and ethical considerations surrounding the use of AI and, in particular, why it can be more problematic than those of conventional safety-related systems. Two important features affecting legal frameworks are highlighted:

1. The potential to change the boundaries of accountability between the human and the machine, and also between design and operation.
2. That the nature of ML is such that transparency and understandability of a machine's decisions, in sufficient time to prevent unintended consequences, may be compromised.

## 11.4 Considerations for ethical practices in safety-related systems

To address the obligation to adhere to the relevant legal requirements within different regulatory regimes, much work is ongoing into developing ethical frameworks for design and use of AI systems. Many of these principles relate to wider societal benefits/drawbacks, however, the key principles for safety-related systems include:

(a) Apply guidance, rules and procedures from relevant acts to ensure ethically mindful development and deployment.

(b) Accountability across all phases of the safety lifecycle (see IEC 61508); it should be borne in mind that AI supply chains are complex and will need careful consideration regarding liability.

(c) Considerations of safe and secure behaviour, including the ability to show that the safety requirements of the safety-related system have been addressed and that the AI will perform as intended (and cannot exhibit unintended dangerous behaviour) under all foreseeable conditions (including misuse, abnormal, failure and malicious attack conditions).

(d) Considerations of how the design of the AI system allows for human oversight, including how sufficient understanding of the impacts of the decision being proposed by the AI system can be sufficiently understood in the time available to prevent unsafe consequences.

(e) If there is ambiguity, seek professional legal advice when adopting AI to ensure it meets their legal obligations.

# References


1. HM Government. *National AI Strategy*. September 2021. https://www.gov.uk/government/publications/national-ai-strategy

2. University College London. *Detecting oesophageal cancer with AI*. July 2021. https://www.ucl.ac.uk/news/2021/jul/detecting-oesophageal-cancer-ai

3. Urie S. The Villanova Law Institute to Address Commercial Sexual Exploitation. *Student Blog Series: Project Artemis: How Microsoft is Using AI Technology to Prevent, Disrupt and Circumvent Traffickers*. 2022. https://cseinstitute.org/project-artemis-how-microsoft-is-using-ai-technology-to-prevent-disrupt-and-circumvent-traffickers/

4. Shaukat K, Luo S, Varadharajan V, Hameed IA, Xu M. *A Survey on Machine Learning Techniques for Cyber Security in the Last Decade*. IEEE Access. 2020. Vol 8, p.222310-222354.

5. Stahl Gasiorowski Criminal Defense. *How the IRS Uses Artificial Intelligence to Detect Tax Evaders*. https://stahlesq.com/irs-artificial-intelligence-detects-tax-evaders/

6. IEC 61508 Parts 1 to 7. *Functional safety of electrical/electronic/programmable electronic safety-related systems*. Edition 2.0. 2010. / BS EN 61508 Parts 1 to 7. *Functional safety of electrical/electronic/programmable electronic safety-related systems*. Edition 2.0. 2010.

7. ISO 26262-1:2018 *Road vehicles - Functional safety. Part 1: Vocabulary.* / BS EN ISO 26262-1:2018 *Road vehicles. Functional Safety - Vocabulary.*

8. IEC TR 62278-4:2016 *Railway applications - Specification and demonstration of reliability, availability, maintainability and safety (RAMS) - Part 4: RAM risk and RAM life cycle aspects.* / BS EN 50126-1:2017 *Railway Applications. The Specification and Demonstration of Reliability, Availability, Maintainability and Safety (RAMS) - Generic RAMS Process.*

9. IEC 62279:2015 *Railway applications - Communication, signalling and processing systems - Software for railway control and protection systems.*

10. IEC 62425:2007 *Railway applications - Communication, signalling and processing systems - Safety related electronic systems for signalling.* / BS EN 50129:2018 *Railway applications. Communication, signalling and processing systems. Safety related electronic systems for signalling.*

11. DO-178C *Software Considerations in Airborne Systems and Equipment Certification*. RTCA Incorporated. 2011.

12. BS EN IEC 62061:2021 *Safety Of machinery. Functional safety of safety-related control systems.* / IEC 62061:2021 *Safety Of machinery - Functional safety of safety-related control systems.*

13. ISO 13849-1:2023 *Safety of machinery - Safety-related parts of control systems - Part 1: General principles for design.*

14. BS EN 61511-2:2017 *Functional safety. Safety instrumented systems for the process industry sector - Guidelines for the application of IEC 61511-1.*

15. ISO/IEC TR 5469:2024 *Artificial intelligence - Functional safety and AI systems.*

16. ISO/CD PAS 8800 *Road Vehicles - Safety and artificial intelligence*. Draft under development.

17. ISO/IEC AWI TS 22440 *Artificial intelligence - Functional safety and AI systems - Requirements*. Approved for development September 2023. Publication planned 2026.

18. International Atomic Energy Agency (IAEA). *Artificial Intelligence for Accelerating Nuclear Applications, Science and Technology*. 2022. https://www.iaea.org/publications/15198/artificial-intelligence-for-accelerating-nuclear-applications-science-and-technology

19. ISO/IEC Guide 51:2014 *Safety aspects - Guidelines for their inclusion in standards.*

20. BS ISO 31000:2018 *Risk management. Guidelines.* / ISO 31000:2018 *Risk management - Guidelines.*

21. BS EN 61882:2016 *Hazard and operability studies (HAZOP studies). Application guide.* / IEC 61882:2016 *Hazard and operability studies (HAZOP studies) - Application guide.*

# References


22. *Systematic Cause Analysis Technique (SCAT).*
https://www.scribd.com/document/401812254/Systematic-Cause-Analysis-Technique-SCAT-doc

23. Leveson NG and Thomas JP. *STPA Handbook*. March 2018.
https://psas.scripts.mit.edu/home/materials/

24. BS ISO 21448:2022 *Road vehicle. Safety of the intended functionality.* / ISO 21448:2022 *Road vehicles - safety of the intended functionality.*

25. Hawkins R, Paterson C, Picardi C, Jia Y, Calinescu R, Habli I. *Guidance on the Assurance of Machine Learning in Autonomous Systems (AMLAS)*. Assuring Autonomy International Programme. University of York. Version 1.1. 2021.

26. *Assurance of Machine Learning for use in Autonomous Systems (AMLAS)*. Assuring Autonomy International Programme. University of York. 2021.
https://www.york.ac.uk/assuring-autonomy/guidance/amlas/

27. Heyn HM, Knauss E, Muhammad AP, Eriksson O, Linder J, Subbiah P, Pradhan SK and Tungal S. *Requirement Engineering Challenges for AI-intense Systems Development*. ArXive.
https://doi.org/10.48550/arXiv.2103.10270

28. Ashmore R, Calinescu R, Paterson C. *Assuring the Machine Learning Lifecycle: Desiderata, Methods, and Challenges*. ArXive. https://arxiv.org/abs/1905.04223

29. Genetic algorithm. Wikipedia. https://en.wikipedia.org/wiki/Genetic_algorithm

30. Marchi JA de, Sharp J, Melrose J, Madahar B, Kurth F, Lange DS, Aktas M, Martinel N, Luotsinen L, Solberg E, Tanik GO. *Robustness of Artificial Intelligence for Hybrid Warfare*. Netherlands Aerospace Centre NLR. 2021.

31. Akerkar RA, Sajja PS. *Knowledge-Based Systems*. Jones & Bartlett Learning. 2010.
https://books.google.co.uk/books/about/Knowledge_Based_Systems.html?id=Tj8r3A-0dZkC&redir_esc=y

32. BS EN ISO/IEC 22989:2023 *Information technology. Artificial intelligence. Artificial intelligence concepts and terminology.* / ISO/IEC 22989:2022 *Information technology - Artificial intelligence - Artificial intelligence concepts and terminology.*

33. IBM. What is supervised learning? https://www.ibm.com/topics/supervised-learning

34. IBM. What is unsupervised learning? https://www.ibm.com/topics/unsupervised-learning

35. University of York. What is reinforcement learning?.
https://online.york.ac.uk/what-is-reinforcement-learning/

36. Banks A and Ashmore R. *Requirements Assurance in Machine Learning*. SafeAI@AAAI. 2019.

37. Floridi L, Holweg M, Taddeo M, Amaya Silva J, Mökander J and Wen Y. *capAI - A Procedure for Conducting Conformity Assessment of AI Systems in Line with the EU Artificial Intelligence Act*. SSRN. 2022.
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4064091

38. National Cyber Security Centre. *Principles for the security of machine learning*. 2022.
https://www.ncsc.gov.uk/files/Principles-for-the-security-of-machine-learning.pdf

39. European Telecommunications Standards Institute (ETSI) Standards Database.
https://www.etsi.org/standards

40. Cybersecurity and Infrastructure Security Agency. EO 14028. *Executive Order on Improving the Nation's Cybersecurity*. 2021.
https://www.cisa.gov/topics/cybersecurity-best-practices/executive-order-improving-nations-cybersecurity#:~:text=Executive%20Order%20(EO)%2014028%2C,adjust%20their%20network%20architectures%20accordingly

41. The Institution of Engineering and Technology. *Code of Practice: Cyber Security and Safety*. 2021.
https://electrical.theiet.org/guidance-and-codes-of-practice/publications-by-category/cyber-security/code-of-practice-cyber-security-and-safety/

## References


42. Bloomfield R, Netkachova K and Stroud R. *Security-Informed Safety: If It's Not Secure, It's Not Safe*. Springer, Berlin, Heidelberg. Vol 8166. 2013.
https://link.springer.com/chapter/10.1007/978-3-642-40894-6_2 / Calinescu R and Di Giandomenico F. *Software Engineering for Resilient Systems*. Springer Cham. 2019
https://link.springer.com/book/10.1007/978-3-030-30856-8

43. Health and Safety Executive. *Reducing risks, protecting people - R2P2*. 2001.
https://www.hse.gov.uk/enforce/expert/r2p2.htm

44. Health and Safety Executive. *ALARP "at a glance"*.
https://www.hse.gov.uk/enforce/expert/alarpglance.htm

45. Health and Safety Executive. *Control Of Major Accident Hazards Regulations 2015 (COMAH)*. 2015.
https://www.hse.gov.uk/comah/background/comah15.htm

46. HM Government. *The Railways and Other Guided Transport Systems (Safety) Regulations 2006*.
https://www.legislation.gov.uk/uksi/2006/599/contents/made

47. HM Government. *The Construction (Design and Management) Regulations 2015*.
https://www.legislation.gov.uk/uksi/2015/51/contents/made

48. European Parliament. *EU AI Act: first regulation on artificial intelligence*.
https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

49. Department for Science, Innovation and Technology. *A pro-innovation approach to AI regulation*. March 2023.
https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1176103/a-pro-innovation-approach-to-ai-regulation-amended-web-ready.pdf

50. Leslie D, Burr C, Aitken M, Cowls J, Katell M and Briggs M. *Artificial intelligence, human rights, democracy, and the rule of law: a primer*. The Council of Europe. 2021.
https://www.turing.ac.uk/news/publications/ai-human-rights-democracy-and-rule-law-primer-prepared-council-europe

51. Center for Strategic & International Studies. *Japan's Approach to AI Regulation and Its Impact on the 2023 G7 Presidency*. February 2023.
https://www.csis.org/analysis/japans-approach-ai-regulation-and-its-impact-2023-g7-presidency

52. The White House. *Blueprint for an AI Bill of Rights*.
https://www.whitehouse.gov/ostp/ai-bill-of-rights/

53. Health and Safety Executive. *Health and Safety at Work etc Act 1974*.
https://www.hse.gov.uk/legislation/hswa.htm

54. Zhou J, Chen F, Berry A, Reed M, Zhang S and Savage S. *A Survey on Ethical Principles of AI and Implementations*. Data Science Institute University of Technology, Sydney. 2020.
https://www.researchgate.net/publication/348263066_A_Survey_on_Ethical_Principles_of_AI_and_Implementations

55. Department for Science, Innovation and Technology, Office for Artificial Intelligence and Centre for Data Ethics and Innovation. *Understanding artificial intelligence ethics and safety*. June 2019.
https://www.gov.uk/guidance/understanding-artificial-intelligence-ethics-and-safety

56. Central Digital and Data Office. *Data Ethics Framework*. September 2020.
https://www.gov.uk/government/publications/data-ethics-framework

# The Application of Artificial Intelligence in Functional Safety

The Application of Artificial Intelligence in Functional Safety focuses on the effective use of artificial intelligence (AI) in safety-related applications, highlighting the importance of underpinning regulation and good practice to embed AI safety. The intention is to highlight the risks associated with AI and the consideration to be given to the various techniques and measures employed during the engineering lifecycle.

This high-level publication aims to provide non-specialist senior managers with appropriate information to support their decision making on the use of AI (particularly in safety-related applications).

What constitutes artificial intelligence has long been debated. Here, in line with the UK Government's National AI Strategy, 2021, we refer to AI as "Machines that perform tasks normally requiring human intelligence, especially when the machines learn from data how to do those tasks." This capacity to learn provides an ability to solve problems that could range from relatively simplistic and specific behaviour, for example, in the approximation of mathematical functions (narrow AI), to the replication of human intelligence (artificial general intelligence (AGI)) and beyond to 'super intelligence'. It should be noted that current technologies have only achieved relatively low levels of narrow AI.

Functional safety is more easily defined. IEC 61508 defines it as relating to that part of the overall equipment under control (EUC) safety that relates to the correct functioning of the electrical/electronic/programmable electronic system safety-related systems.

This document focuses on 10 key 'pillars' that highlight the risks of using AI in such systems. It outlines the additional considerations required in the engineering processes and provides guidance on building assurance cases for AI in safety related applications. It also addresses the fundamental differences between traditional and AI software.

**The Institution of Engineering and Technology**
Futures Place
Kings Way
Stevenage
Herts
SG1 2UA

**theiet.org/electrical**