| | | | |
|---|---|---|---|
| **Document filename:** Healthcare Supplementary Guidance for AMLAS | | | |
| **Directorate / Programme** | Clinical Safety | **Project** | AAIP - SAFR |
| **Document Reference** | | | |
| **Project Manager** | Sean White | **Status** | Final |
| **Owner** | Shakir Laher | **Version** | 1.0 |
| **Author** | Shakir Laher | **Version issue date** | 05.12.2022 |

# Healthcare Supplementary Guidance for AMLAS

# Document Management

## Revision History

| Version | Date | Summary of Changes |
|---|---|---|
| 0.1 | 26.05.2022 | Initial draft |
| 0.2 | 01.07.2022 | SW comments actioned |
| 0.3 | 26.07.2022 | CB & IH comments actioned |
| 1.0 | 05.12.2022 | Director Approval |
| | | |
| | | |

## Reviewers

This document must be reviewed by the following people:

| Reviewer name | Title / Responsibility | Date | Version |
|---|---|---|---|
| Sean White | Principal Safety Engineer | 01.07.22 | 0.1 |
| Carla Brackstone | Partnerships Manager | 25.07.22 | 0.2 |
| Professor Ibrahim Habli | AAIP Technical Lead | 28.07.22 | 0.2 |

## Approved by

This document must be approved by the following people:

| Name | Signature | Title | Date | Version |
|---|---|---|---|---|
| Dr Manpreet Pujara | | Director of Patient Safety | 05.12.2022 | 1.0 |
| | | | | |

## Glossary of Terms

| Term / Abbreviation | What it stands for |
|---|---|
| | |
| SAFR | Safety Assurance Framework for Machine Learning in the Healthcare Domain |
| AMLAS | Assurance of Machine Learning for use in Autonomous Systems |
| ML | Machine Learning |
| HCP | Healthcare Professional |
| SME | Subject Matter Expert |
| PHA | Preliminary Hazard Analysis |
| HAZOP | Hazards and Operability Analysis |
| HAZID | HAZard IDentification |
| SWIFT | Structured What If Technique (SWIFT) |
| SHARD | Software Hazard Analysis and Resolution in Design (SHARD) |

| | |
|---|---|
| FRAM | Functional Resonance Analysis Method (FRAM) |

**Document Control:**

The controlled copy of this document is maintained in the NHS Digital corporate network. Any copies of this document held outside of that area, in whatever format (e.g. paper, email attachment), are considered to have passed out of control and should be checked for currency and validity.

# Contents

# 1 Introduction

## 1.1 Purpose of document

The Assurance of Machine Learning for use in Autonomous Systems (AMLAS) is a methodology for establishing justified confidence in the safety of Machine Learning (ML) components within a wider system and context [1]. It comprises systematic and structured processes, covering six ML life cycle stages, accompanied by argument patterns that can be followed to establish a ML safety case. AMLAS is domain independent and represents state-of-the-art in ML safety assurance, building on best practice in safety-critical systems engineering. Therefore, this supplementary guidance provides healthcare specific considerations that should be reviewed in parallel when applying AMLAS.

## 1.2 Using this Document

Manufacturers should read this document after they have familiarised themselves with AMLAS. The guidance is structured into five key themes to be considered throughout application of AMLAS. Each theme is expanded and explained through a series of sub-topics/points. Every effort has been made to reference the guidance to a specific AMLAS stage's process or argument pattern. Where this has not been possible, it should be understood as that sub-topic/point needing to be considered throughout.

A table at the beginning of each theme provides clarity by offering a summary of actions a manufacturer needs to take for each sub-topic/point by referencing them to a specific area of importance, or if the sub-topic/point needs to be considered throughout the application of AMLAS, it is denoted as 'General Consideration'.

This document does not replicate any of the AMLAS content and therefore all that applies to AMLAS in terms of scope, intended users and definitions/terms is applicable to this guidance.

# 2 Guidance Section

## 2.1 Whole System Safety – Identify Full Range of Hazards to Support Derivation of System Safety Requirements

| Sub-topic/point | Actions |
|---|---|
| Identifying full range of hazards | Establish Artefact A Stage 1 |

Healthcare clinical pathway systems are intrinsically socio-technical and are constructed of three core components, technology, healthcare professional (HCP) and information. Each component has a specialised role, as will be the case for any ML component. However, these core components must all come together to achieve an overarching goal of providing healthcare services safely.
AMLAS adheres to a well-established safety engineering concept of system safety, whereby safety of a system component is only meaningful if its contribution to safety is considered from a system-level within the intended context. For this reason, prior to applying the stages of AMLAS there is a pre-requisite for system safety requirements [**Stage 1: Artefact A**] to be established which should be derived by initially conducting system safety analyses that identify system-level hazards.

**Considerations for identifying full range of hazards to establish system safety requirements**

1. **Technique selection:** selecting analyses techniques that are fit for purpose will enhance the chances of ***identifying full range of hazards*** that are caused in healthcare by technology, human factors, or information flows. It is feasible, more than one technique may need to be utilised to identify the hazards. Table 1 offers existing analysis techniques to capture the hazards, although it is by no means exhaustive.
2. **HCP Inclusion:** prior to implementing AMLAS, identification of credible system level hazards and corresponding safety requirements is vital due to their cascading dependency in each AMLAS stage. This identification is unlikely to happen correctly, and in a timely manner, if only those who have an engineering or safety role are consulted. Therefore, explicit inclusion of HCPs should be considered. Their expertise, knowledge and skills obtained from being embedded in clinical pathways will enrich output.
3. **Analyses scope**: all analyses must be scoped within a ***boundary of responsibility*** to identify credible hazards (i.e., potentially lead to patient harm) from which safety requirements may be derived. To this aim, a clear understanding of the clinical pathway is required to realise in context each component's influence from a safety perspective. Graphical process flow charts can be one of many approaches to achieving this objective.

4. **Establish requirements**: functional safety requirements (i.e., requirements established to ensure a product operates correctly to maintain the overall safety [2]) are the obvious choice when establishing safety requirements for technology. However, full consideration must be given to the safety requirements associated with human factors, such as, situational awareness, automation bias and human-machine teaming, to mention a few [3]. Furthermore, ML has brought into scope attributes typically associated with humans from which now there is a need for safety requirements to be derived. A few of these attributes are, interpretability, explainability and fairness, which all fall under the umbrella term, transparency.
5. **Safety requirements allocation to ML**: once derivation of system safety requirements is complete, they will become **Artefact A** as labelled in AMLAS stage 1. Subsequent AMLAS activities will expect allocation of ML safety requirements derived from the system safety requirements, which the ML must then satisfy through the ML development lifecycle, as prescribed by AMLAS stages 2-6.

| Technique | Description |
|---|---|
| Preliminary Hazard Analysis (PHA) | An iterative technique used early in the life cycle to identify broad system hazards to inform safety design criteria and requirements [4] |
| Hazards and Operability Analysis (HAZOP) | Studies how work process flows may deviate from its design intent. It systematically examines the potential for failure of the individual and consequential effects for the whole system. [5] |
| HAZard IDentification (HazID) | This technique makes use of guidewords (e.g., incorrect, omission, late, etc) related to the system to identify hazards [6] |
| Structured What If Technique (SWIFT) | High-level risk identification technique conducted through the use of guidewords (e.g., 'what if …' or 'how could …', etc) [7] |
| Software Hazard Analysis and Resolution in Design (SHARD) | A variant of HAZOP, particularly suited to identifying hazards for computer-based systems where information flows through systems. Applied in a structured manner through the use of guidewords (incorrect, early, late, commission and omission) [8] |
| Functional Resonance Analysis Method (FRAM) | A system-based analysis method used to model and understand interactions between technologies and humans that are part of a socio-technical system [9] |

**Table 1 - Example of System Level Hazard Analysis Techniques**

## 2.2 Explicit HCP Inclusion

| Sub-topic/point | Actions |
|---|---|
| HCP as a SME | Review Table 2 – HCP input mapped to AMLAS |

**HCP as a SME**

HCPs have expertise, knowledge and skills acquired from clinical practice. As subject matter experts (SMEs) they should be included throughout the safety work. It may not be the same HCP that participates in activities and a safety lead (e.g., clinical safety officer), or the multi-disciplinary team will need to decide, who, and at what stage they are needed. Table 2 provides guidance as to where in AMLAS their input would be beneficial, although this is not prescriptive and professional judgement should be exercised:

| AMLAS Stage | Artefact/Activity | Benefit of HCP Inclusion |
|---|---|---|
| 1 | Artefact [B]: Description of Operating Environment | Describe with greater accuracy the '*operational environment*' which translates in healthcare as, the clinical pathway. |
| 1 | Activity 1: Define the Assurance Scope for the ML Component | Will be able to anticipate safety scope in context coupled to specific set of use cases. |
| 2 | Activity 3: Develop ML Safety Requirements | Identify and assign safety requirements bound to a clinical context, e.g., explainability of ML output for HCPs. |
| 3 | Activity 8: Validate ML Data | Appraise whether the data are relevant, complete, accurate & balanced to satisfy safety requirements from a clinical perspective. |
| 6 | Activity 16: Test the Integration | Appraise whether safety requirements are satisfied through clinical practice. |
| 6 | Artefact [EE]: Operational Scenarios | Contribute towards describing '*operational scenarios*' which translates to specific healthcare use cases. |

**Table 2 – HCP input mapped to AMLAS**

## 2.3 Performance Metrics vs Soft Constraints

| Sub-topic/point | Actions |
|---|---|
| High performance ≠ to safe performance | General Consideration |
| Metric Selection | General Consideration |
| Soft Constraints | Extend S2.2 Stage 2 |

**High performance ≠ to safe performance**

ML Models are able to generalise with very high levels of accuracy on narrow well defined tasks. This provides utility to identify healthcare conditions from their well-understood identification routes (typical pathology). However, healthcare conditions often have lesser known variants collectively known as rare conditions. Caution should be exercised to understand the ability of the model to identify these conditions and how they have been included in the dataset. If there is insufficient data for the rare conditions, the high performance metric is unlikely to be accurate in its claim to identifying them. Therefore, it should be fully understood why specific ML performance metrics have been chosen to contribute towards safety requirements. It will invariably require a collection of metrics to deal with the array of variations that are present in healthcare due to patient demographics and types of conditions.
Inclusion of edge cases (i.e., occur at the extreme ends of the spectrum or are unforeseen) within the dataset will assist towards the above.

**Metric Selection**

When deciding on safety requirements that are linked explicitly to performance metrics it is important to justify why a specific metric has been selected for clinical and patient safety, e.g., appropriate balance of sensitivity vs specificity for a specific clinical use case. There is no one best fit answer and trade-offs will need to be made to achieve the right balance to satisfy safety.

**Soft Constraints**

Contrary to performance metrics, which are hard constraints due to their quantitative nature, soft constraints (i.e., a property that should be satisfied, e.g., trust) will need careful thought and planning with foresight to be identified as a safety requirement and then satisfied based on set criteria which may be largely qualitative. This poses a challenge, as accurate identification requires a need for ML engineers, HCPs and safety practitioners to collaborate early in the safety requirements elicitation stages (AMLAS stages 1 & 2). It is recommended efforts are made for this to happen at the appropriate stages rather than retrospectively.

A further consideration here is being aware that definitions associated with soft constraints are ambiguous and depend on who is defining the term and for which end-user. A recommendation here is for ***interpretability*** and ***explainability*** to be explicitly considered on a case-by-case basis, although thinking should not be limited to them alone. AMLAS allows **S2.2** ('Argument over satisfaction of different types of ML safety requirements') within stage 2 argument pattern to be extended to include any additional safety requirements beyond the prescribed performance and robustness. **S2.2** description (p.17 AMLAS) provides further guidance.

## 2.4 Data Management

| Sub-topic/point | Actions |
|---|---|
| Data Drift | • Assign a data drift safety requirement to the ML at stage 2<br>• Satisfy the requirement as part of S6.2 Stage 6 |
| Bias | Consider bias explicitly at S3.2 Stage 3 |

**Data Drift**

Offline supervised ML models trained on data during their development will provide a level of performance at a point in time, i.e., completion of the training cycle. On live use, this performance may drift as the model begins to ingest operational data which it may not have been trained on. Therefore, if a model encounters previously unseen data it should still be able to make predictions or flag a safety warning. This is a likely scenario due to the diverse populations that healthcare systems serve, and as healthcare conditions mutate and evolve. This can lead to data drift (change in distribution between the training and live use data). A data drift safety requirement will need to be assigned to the ML and satisfied during the deployment stage (6) of AMLAS as part of **S6.4**.

**Bias**

An ML model at its core is data driven. Therefore, any biases present in the data sets will be learned and could lead to unfair output by the model. However, to completely eradicate biases from data is almost impossible and the goal should be to identify and eradicate to such an extent that the safety of patients is preserved.
Recommendation here is to train models with data that have specific bias-centric (e.g., ethnicities) safety requirements assigned, specifically to prevent unfair discrimination against patient sub-populations based on demographics.

AMLAS provides four attributes in the form of, relevant, accurate, complete and balance (**S3.2, Stage 3 Argument**), to apply as tests to data sets used. The word bias is not explicit in the four attributes and should therefore be a consideration throughout application of the four attributes. Manufacturers should be particularly aware of the misconception that acquiring historic data sets from the target healthcare pathway as being unbiased. Health datasets are created by technology and differing HCPs via ever-changing hardware, software, infrastructure, architecture, standard operating protocols, fluid patient populations and evolving/emergent health conditions. These intersecting combinatorial factors will have left biases that will require close inspection to firstly identify, and then minimise or eradicate.

## 2.5 Development Environment to Live Deployment

| Sub-topic/point | Actions |
|---|---|
| Distribute responsibilities | General consideration within Argument Pattern Stage 6 |
| Define the intended use | Include explicitly within Artefact D Stage 1 & Artefact V Stage 4 |
| Routine monitoring | Consider as part of G6.3 Argument pattern Stage 6 |
| Modification | Consider as part of G6.8 Argument pattern Stage 6 |

ML models built to be embedded into a healthcare pathway will require collaborative working between a manufacturer and the adopter (e.g., healthcare organisation). This is imperative due to the adaptive nature of healthcare services and practices. The manufacturer will have strong skills and knowledge from a technical perspective however, it is the adopter that will be utilising the technology on a daily basis and will bring the specialist knowledge of the clinical pathway. Overlooking specialist user knowledge can lead to difficulties in integrating the technology within an existing socio-technical system and continuing to monitor its safe operation. For these reasons, when working through the deployment stage (6) of AMLAS the following should be considered:

- **Distribute responsibilities clearly** between the manufacturer and adopter for the safe integration of the ML into the clinical pathway to ensure safety requirements are still applicable in a live setting. If a manufacturer asserts that they can integrate and monitor safe operation without explicit input from the adopter, clear evidence of this assertion should be sought. This point is applicable throughout the stage 6 argument pattern.
- **Define the intended use** of the ML component from the outset (**Artefact D**) and subsequent ML model (**Artefact V**), including limitations. This will ensure the functional scope is not extended causing an unsafe state to arise.
- **Routine monitoring** - as the manufacturer will not aways be present on site, be clear in how the ML model will be monitored between the manufacturer and adopter to ensure safety requirements are being satisfied. Both stakeholders will need to be vigilant as degradation of a system can be difficult to identify, although in the event of this fail safes and contingencies must be ready to deploy. This should be considered as part of **G6.3**.
- **Modification** - if any changes are proposed for the ML, how they will be implemented and who will be responsible for the change management to ensure safety requirements continue to be satisfied must be agreed. This should be considered at **G6.8**.

# 3 References


[1] R. Hawkins et al., *Guidance on the Assurance of Machine Learning in Autonomous Systems (AMLAS)*, 2021.
https://www.york.ac.uk/assuring-autonomy/guidance/amlas/

[2] International Electrotechnical Commission, *IEC 61508:2010 CMV Commented Version*, 2010, IEC. https://webstore.iec.ch/publication/22273

[3] Sujan, M et al., *Human factors challenges for the safe use of artificial intelligence in patient care.* 2019. BMJ health & care informatics 26, no. 1.

[4] N. G. Leveson, *Safeware: system safety and computers*. New York, NY, USA: Association for Computing Machinery, 1995. Pgs. 153, 341.

[5] Dunjo, J., et al., *Hazard and Operability (HAZOP) analysis. A literature review*. 2010. Journal of Hazardous Material. Vol. 173, pgs 19-32.

[6] Palm. M., et al., *Risk analyses in the intersection between patient and workplace safety: A case study of hazids in para-clinical supporting systems in specialized health care*. 2020. SAGE Open Medicine.

[7] Card, A. J., *Beyond FMEA: The Structured What-If Technique (SWIFT)*. Journal of Healthcare Risk Management, 2013

[8] Pumfrey, D.J. *The principled Design of Computer System Safety Analyses*. PhD Thesis. 1999. University of York

[9] Hollnagel, E. *FRAM: The Functional Resonance Analysis Method*. 2016 London: CRC Press. https://doi.org/10.1201/9781315255071