Body of Knowledge - Structure and Scope

Richard Hawkins

April 2019

UNIVERSITY of York

ASSURING AUTONOMY
INTERNATIONAL PROGRAMME

Table of contents

1. Introduction

The scope of the Assuring Autonomy International Programme (AAIP), and therefore the Body of Knowledge (BoK) is huge. It must be cross-domain, cross-technology and cross-application and cover all aspects of assurance and regulation of Robotics and Autonomous Systems (RAS). In addition, it must present information which is accessible and useful to a range of different stakeholders. Some of the knowledge will inevitably be specific in nature, e.g. to domain or technology, however where possible the guiding principles are that the guidance should:

- Be as general as possible, only as specific as necessary, but providing domain-specific guidance where required;
- Use established assurance approaches.

As part of this it is our intention to indicate how traditional, established safety engineering and assurance methods can be used where possible. However, the BoK will not provide guidance on how to do this, other than to show how to apply these existing approaches to autonomous systems (the BoK is not going to be a manual on general safety engineering). In particular we aim to identify, and focus on the key areas where assurance for RAS is particularly different to standard safety assurance practice and challenging conceptually or practically.

This document provides a proposed structure for the BoK that should enable us to meet the above specification. Developing the BoK is a difficult endeavour, and it is inevitable that the structure and content of the BoK will evolve over time, thus this should be viewed as an initial structure which will be revised as necessary. As well as describing the structure, information is also provided on the scope of what will be covered under each part of that structure. Again, this is subject to change and it is anticipated that the BoK will evolve during the life of the AAIP, and beyond.

The BoK will cover 4 main areas, as set out below. These are chosen as they support core principles for developing, assuring and regulating RAS:

- Definition of required behaviour - Defining what it means for the RAS to be ‘safe’;
- Implementing an RAS to provide the required behaviour - Demonstrating the sufficiency of the implementation;
- Understanding and controlling deviations from required behaviour – Identifying and controlling sources of deviation;
- Gaining approval for operation of RAS – Gaining approval for operation in the specified environment from the relevant regulatory authority.

The BoK must include consideration of how these are achieved through life, not just at design-time; this is necessary for all classes of system, but particularly important where RAS learn in operation, hence their behaviour is not ‘fixed’ at design-time. It must also consider

the cases where there are multiple interacting systems, often referred to as a System of Systems (SoS).

Below, a hierarchical structure for the BoK is established. Within this structure:

- Each heading represents an assurance or regulatory consideration on which guidance will be provided (items highlighted in red). Each 'entry' will consist of three main parts:
    - A definition of ***assurance objectives*** relating to that area of consideration. The assurance objectives are things that must be demonstrated when putting an RAS into operation.
    - A ***contextual description*** that provides further information and rationale for the assurance objectives.
    - Details on ***approaches for demonstration*** of meeting the assurance objectives. There will often be multiple alternative strategies for demonstration. These alternative strategies may reflect alternatives in the state-of-practice, or reflect different approaches required in different domains or with different technologies. Where appropriate reference will be provided to further public domain information.

This structure is defined in section 2 using the following format:

Assurance consideration X

Assurance (or Regulatory) Objectives

Contextual Description

Approaches for Demonstration

It is anticipated that the objectives and contextual description will be relatively stable, but that the approaches to demonstration will evolve as the state-of-the-art matures, e.g. new approaches are identified for verifying machine learning (ML). The Regulatory Objectives draw on the Assurance objectives, but address the considerations in approving the initial and ongoing operation of a RAS.

Section 3 provides a graphical summary of the BoK structure, and section 4 provides a template assurance case (for a single RAS not an SoS) that reflects the considerations in the BoK. This is cross-linked so that the information in the BoK can be seen as guidance on how to demonstrate the claims in the assurance case.

At this stage there are no 'hard and fast' rules for the scale of entries in the BoK. The intent is that the definition of assurance objectives are succinct and that they should require little technical knowledge to understand. The context description is typically one or two paragraphs, again intended to be readily accessible to stakeholders. The description on

approaches to demonstration is likely to be more technical and more variable in length. If the approach is well-understood, the BoK entry may be primarily a set of references to the literature, with an indication of how to map the approach to the stated objectives. For material developed by the AAIP and not (yet) published the entry may be much more extensive. However, where possible an accessible summary will be provided. In later versions of the BoK it may well be desirable to include summaries for different classes of stakeholder, e.g. lawyers and researchers; at this stage the focus is technical.

2. BoK structure and scope

Rationale for including the individual considerations is reflected in the contextual description. As indicated in the introduction, the BoK structure is presented as a hierarchy. However, the problem of assuring and regulating RAS is complex, and the elements of the BoK are interdependent. For simplicity, cross-references are not shown explicitly, although they are implied in some cases, e.g. the role of simulation in validation safety requirements. However, it should be assumed that all elements of the BoK structure need to be addressed, to some degree, to have a complete approach to assurance and regulation.

1 Defining required behaviour

Assurance Objective: Define how the RAS must behave in order to be sufficiently safe.

Contextual Description: The primary objective for safety assurance of any system is to demonstrate that the system's behaviour is sufficiently safe throughout its life. The first stage of this is to understand, and to specify, what is considered to be sufficiently safe behaviour for the system. This will include defining what the RAS must do, as well as what it must not do, in order to be considered to be sufficiently safe. In order to define this appropriately for an RAS, there are a number of objectives that must be satisfied, as described below.

1.1 Identifying Hazards

Assurance Objective: Identify the hazards associated with the operation of the RAS.

Contextual Description: It is not possible to define safe behaviour without first identifying all of the hazards associated with the RAS operation. Although it is not possible to prove that all hazards have been correctly identified, it is important to demonstrate there is sufficient confidence in the completeness and correctness of the hazards for the defined operation of the RAS through the approach taken to hazard identification.

Approaches for Demonstration: *In many cases standard hazard identification techniques such as structured brainstorms or checklists will be applicable, particularly*

in a closed environment where the RAS is replacing a human in well understood tasks. More guidance will be required in open environments and where the RAS is introducing novel behaviour.

1.1.1 Defining system scope

Assurance Objective: Define the system boundary for the RAS.

Contextual Description: It is important to be clear about what the system is for which assurance is being provided as anything that falls outside of that system definition will not be considered during the assurance process, with responsibility for its assurance lying elsewhere. It is particularly important for RAS whose behaviour is distributed amongst a number of entities, or where humans and machines are interacting that the scope of the system under consideration is clearly defined to ensure the completeness of the assurance activities. Where multiple interacting systems are used, a decision is required on whether to consider systems as independent entities, or to take a holistic SoS view.

Approaches for Demonstration: *TBD*

1.1.2 Defining the operating environment

Assurance Objective: Define the environment in which the RAS will operate.

Assurance Objective: Define assumptions relating to the environment in which the RAS will operate.

Contextual Description: The environment in which the RAS operates will determine the hazards associated with the RAS. The definition of the environment must be correct, and remain correct throughout the operational life of the RAS to ensure the definition of the hazards and hazardous behaviour for the RAS remain valid. Defining the environment may require assumptions to be made, and these assumptions must be included in the definition of the environment. The continued validity of the operating environment definition and assumptions may need to be monitored through-life, this is addressed as a separate consideration.

Approaches for Demonstration: *TBD*

1.1.3 Defining operating scenarios

Assurance Objective: Define the operating scenarios of the RAS.

Assurance Objective: Define assumptions regarding the operating scenarios of the RAS.

Contextual Description: To fully understand the hazards associated with the RAS, it is necessary to understand how the RAS is expected to operate. This can be defined as a set of scenarios for the operation of the RAS. In general, a RAS will only operate in some scenarios that would be possible given the operating environment[1]. The operation of the RAS must correspond to the defined scenarios throughout the operational life of the RAS to ensure that continued validity of the hazards and hazardous behaviour defined for the RAS. It will rarely, if ever, be possible to completely define all the operating scenarios for the RAS. The challenge is to identify an appropriate level of detail to understand the hazards sufficiently. Defining the operating scenarios may require assumptions to be made, and these assumptions must be included in the definition. The continued validity of the operating scenarios defined, and assumptions made, may need to be monitored through life, this is addressed as a separate consideration.

Approaches for Demonstration: *TBD*

1.2 Identifying hazardous system behaviour

Assurance Objective: Identify how the RAS could bring about hazards given its defined operation and environment.

Contextual Description: Having identified system hazards, the ways in which the system may bring about those hazards must be determined. This will require consideration of both nominal and deviant behaviour of the system. An important consideration is that unusual or unexpected behaviour of the RAS, although not necessarily directly hazardous to the RAS itself, may provoke behaviour in another system or human that is potentially hazardous.

Approaches for Demonstration:

Standard techniques such as Functional Failure Analysis (FFA) and HAZOP may be used here, but for RAS additional guidance on their application may be required (including potentially additional guidewords etc). Alternative techniques such as simulation may also be required in order to fully explore the behaviour of the system. Possible security

[1] For example, for an autonomous (self-driving) car the operating environment might be the City of York, inside the outer ring road. However, the scenarios might limit autonomous operation to certain weather conditions and particular times of day, and exclude all pedestrianised areas, even though vehicular access to such areas is permitted at certain times of day.

attack scenarios should also be considered to identify if these could result in system hazards.

1.2.1 Considering Human/Machine interactions

Assurance Objective: Identify how interactions between humans and the RAS could bring about hazards.

Assurance Objective: Identify training requirements for humans who interact with the RAS.

Contextual Description: As part of identifying how the system may bring about hazards it is important to consider the necessary interactions between the RAS and the human (whether that is a system operator, or a third-party in the same environment as the RAS). A particular concern for RAS may often lie in the handover of control between the human and the machine, and the level of situational awareness that this might require on the part of the human. As well as identifying where such interactions may lead to hazardous behaviour, it is also necessary to consider the ways in which humans may need to be trained in order to interact safely with RAS.

Approaches for Demonstration: *TBD*

1.3 Defining safety requirements

Assurance Objective: Define safety requirements for the RAS sufficient to ensure safe behaviour.

Contextual Description: The safety requirements specify what the RAS must achieve during operation in order to be considered sufficiently safe. The high-level safety requirements can be largely defined based on the understanding of how hazards may arise for the RAS. These high-level safety requirements must later be further refined to more detailed requirements.

Approaches for Demonstration:

In most cases it is anticipated that essentially standard approaches to requirements specification will be adopted.

1.3.1 Validation of safety requirements

Assurance Objective: Demonstrate that the defined safety requirements are sufficient to ensure safe behaviour of the RAS.

Contextual Description: It must be demonstrated that the defined safety requirements are a sufficient specification of safe behaviour for the RAS within its defined operating environment and scenarios of use.

Approaches for Demonstration:

It may be that different stakeholders have different views on what is considered sufficiently safe. For this reason it is often important to involve a range of stakeholders as part of the validation process.

Simulation can also be used in order to illustrate the specified behaviour.

1.4 Impact of security on safety

Assurance Objective: Identify the potential impact of security threats on the safety of the RAS at all stages of the assurance process.

Contextual Description: Security is concerned with the prevention of loss arising from malicious causes. Security often focusses on loss of data or financial loss, and as such security assurance in general is not the focus of the BoK. However, security attacks on a system may also impact the safety of that system by giving rise to hazards. It is therefore important that the contribution of security is considered as part of the safety assurance process.

All systems are vulnerable to security attacks to some extent, however the nature of many RAS makes them particularly vulnerable, and may introduce a number of unique security challenges. It is important that the effects of security are considered throughout the safety assurance process.

***Approaches for Demonstration**:*

General guidance on the nature of security threats for RAS. Specific guidance on the impact of security threats and how they should be considered will be provided against the particular assurance objectives.

2 Implementation of an RAS to provide the required behaviour

Assurance Objective: Implement an RAS that demonstrably satisfies the defined safety requirements.

Contextual Description: Having defined how the RAS must behave in order to be sufficiently safe, it is then necessary to implement the RAS such that it provides that behaviour throughout its life, and to provide sufficient evidence that this has been achieved. In order to define this appropriately for an RAS, there are a number of objectives that must be satisfied, as described below. System requirements are implemented through a process of

architecture and design decomposition. Although this process may vary enormously for different systems and domains, it is generally possible to consider an RAS in terms of an agent model consisting of the following elements:

- Sensing
- Understanding
- Deciding
- Acting

The relationship between these elements is indicated in Figure 1. Each of these elements may be further decomposed into components that implement that aspect of the RAS behaviour. Note that not all components need be part of the RAS itself — they may be part of infrastructure provided externally, e.g. an autonomous car may perform some Sensing by receiving information from roadside beacons.).

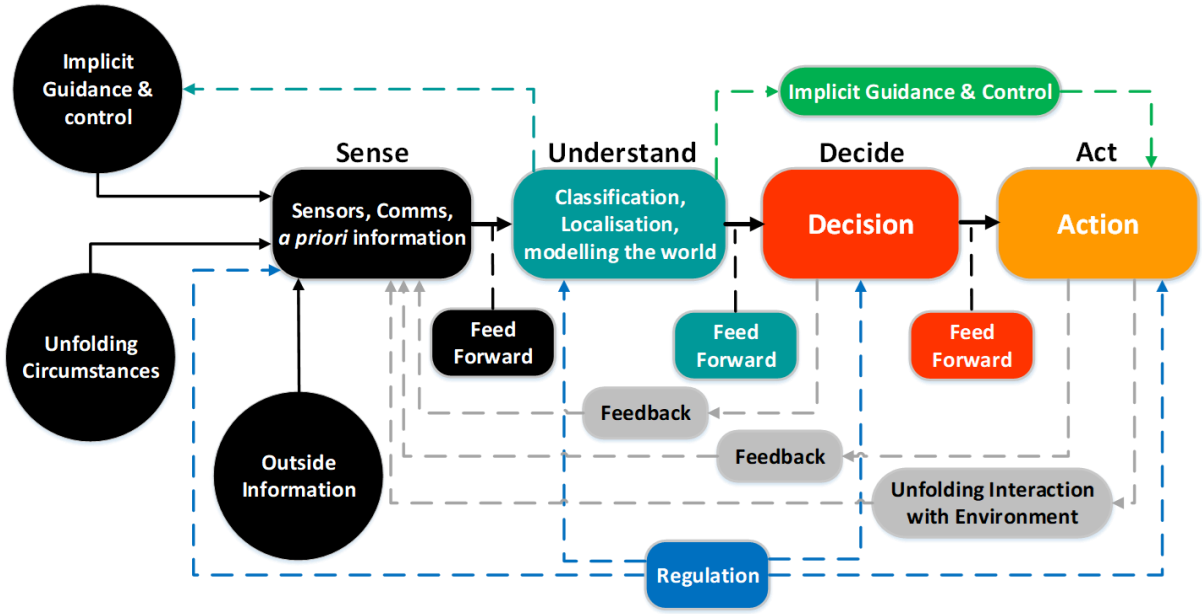


Figure 1 - SUDA agent model

2.1 System-level verification

Assurance Objective: Provide evidence that system-level behaviour satisfies the defined safety requirements.

Contextual Description: As part of demonstrating that the required behaviour is achieved, the performance of the system as a whole should be assessed against the safety requirements. This will provide evidence that may be used as part of a safety justification for the RAS.

Approaches for Demonstration: *Predominantly this objective would be demonstrated through testing, either in the real world or through use of simulation, taking account of the strengths and weaknesses of such approaches.*

An alternative or complementary approach is to generate evidence using formal verification. Guidance will be provided on the use of formal techniques and the advantages and challenges of doing so.

2.2 Implementation of SUDA elements

2.2.1 Defining requirements for SUDA elements

Assurance Objective: Define safety requirements for each element of the RAS architecture sufficient to ensure safe behaviour of that element.

Contextual Description: The safety requirements of the RAS must be allocated, apportioned and interpreted for each element of the RAS architecture (Sensing, Understanding, Deciding, Acting and Infrastructure). The safety requirements must define what each element must achieve if the safety requirements defined for the RAS as a whole are to be satisfied. The safety requirements for each element must take account of the defined operating scenarios, as well as the environmental assumptions that have been made (for example, whether the required behaviour needs to be achieved at night or in heavy rain).

Approaches for Demonstration:

The way in which this objective is demonstrated may be different for different elements of the architecture as defined in the following sub-sections.

2.2.1.1 Defining Sensing requirements

Approaches for Demonstration: *TBD*

2.2.1.2 Defining Understanding requirements

Approaches for Demonstration: *TBD*

2.2.1.3 Defining Deciding requirements

Approaches for Demonstration: *TBD*

2.2.1.4 Defining Acting requirements

Approaches for Demonstration: *TBD*

2.2.1.5 Defining Infrastructure requirements

Approaches for Demonstration: *TBD*

2.2.1.6 Validation of requirements for SUDA elements

Assurance Objective: Demonstrate that the safety requirements defined for each element of the RAS architecture are sufficient to ensure the safe behaviour of that element.

Contextual Description: It must be demonstrated that the defined safety requirements for each element are a sufficient specification of what that element must or must not do if the safety requirements defined for the RAS as a whole are to be satisfied, in the defined operating context and scenarios of use.

Approaches for Demonstration: *TBD*

2.2.2 Defining requirements on components

Each of the SUDA elements may be implemented by a number of different components. For example, multiple components of different types might be used to provide the overall sensing capability of the RAS. It is important where this is done that the assurance of the individual components is considered.

Assurance Objective: Define safety requirements for each component that are sufficient to ensure safe behaviour of that component.

Contextual Description: Once the requirements each element of the RAS architecture are known and decisions have been made as to the components that will be used to implement this, more specific safety requirements on each of those components must be defined. These requirements must define what each component has to achieve if the safety requirements defined for that element of the RAS architecture are to be satisfied.

Approaches for Demonstration:

The approaches for demonstrating this objective will inevitably be technology specific, since they involve an understanding of the capabilities of particular

components. Guidance may be provided on the advantages and limitations of different types of components in different domain applications as defined in the following sub-sections.

2.2.2.1 Defining requirements on 'Sensing' components

Approaches for Demonstration: *TBD*

2.2.2.2 Defining requirements on 'Understanding' components

Approaches for Demonstration: *TBD*

2.2.2.3 Defining requirements on 'Deciding' components

Approaches for Demonstration: *TBD*

2.2.2.4 Defining requirements on 'Acting' components

Approaches for Demonstration: *TBD*

2.2.2.5 Defining requirements on Infrastructure components

Approaches for Demonstration: *TBD*

2.2.2.6 Validation of requirements on components

Assurance Objective: Demonstrate that the safety requirements defined for each component are sufficient to ensure the safe behaviour of that component.

Contextual Description: It must be demonstrated that the defined safety requirements for each component are a sufficient specification of what that component must achieve if the safety requirements defined for the relevant element of the RAS architecture are to be satisfied.

Approaches for Demonstration: *TBD*

2.2.3 Controlling interactions between components

Assurance Objective: Identify how interactions between components may give rise to unsafe behaviour.

Assurance Objective: Manage interactions between components to ensure they do not result in unsafe behaviour.

Contextual Description: Multiple components will often be required in order to implement the safety requirements. Although individual components may meet their requirements, it may still be possible for unsafe behaviour to emerge due to the interactions between those components. It is therefore required to provide sufficient confidence that potentially unsafe interactions between components have been identified and mitigated. Mitigation may require additional safety requirements to be derived and implemented.

Approaches for Demonstration: *TBD*

2.2.4 Verification of requirements for SUDA elements

Assurance Objective: Demonstrate that the safety requirements defined for each element of the RAS architecture are satisfied.

Contextual Description: Evidence must be generated to provide sufficient confidence that the defined safety requirements are satisfied by the implementation of each element.

Approaches for Demonstration: *TBD*

Many approaches for demonstrating this objective will be standard verification approaches, however there may be areas such as the role of simulation in testing that are particular to RAS.

ML may be used as part of the implementation of some elements. Approaches for verification of machine-learnt components are considered under a separate objective.

The way in which these objectives are demonstrated may be different for different elements of the architecture as defined in the following sub-sections.

2.2.4.1 Verification of Sensing requirements

Approaches for Demonstration: *TBD*

2.2.4.2 Verification of Understanding requirements

Approaches for Demonstration: *TBD*

2.2.4.3 Verification of Deciding requirements

Approaches for Demonstration: *TBD*

2.2.4.4 Verification of Acting requirements

Approaches for Demonstration: *TBD*

2.2.4.5 Verification of Infrastructure requirements

Approaches for Demonstration: *TBD*

2.3 Implementing requirements using ML

Assurance Objective: Provide a ML implementation that meets the defined safety requirements.

Contextual Description: ML may be used as part of the implementation of any of the 'SUDA' functions, but in practice is most likely for Understanding and Deciding. Where ML is used as part of the implementation, it is necessary to ensure that the implementation satisfies the allocated safety requirements. Different types of machine learning technology may be adopted including neural networks, Bayesian networks, random forests and reinforcement learning, and the implications that technology choices may have on assurance must be considered.

This objective is achieved through the consideration of three sub-objectives as described below. These sub-objectives reflect the main elements of an ML process as shown in Figure 2.

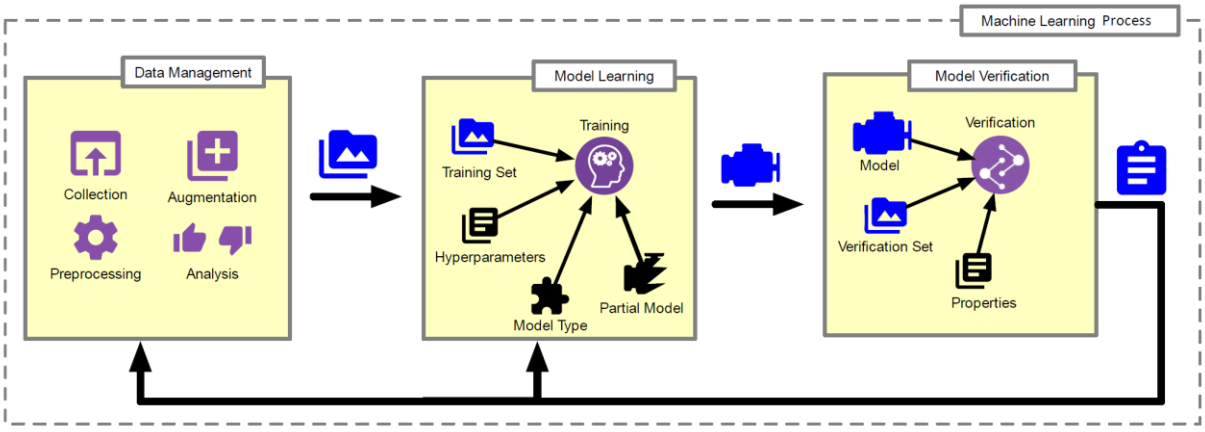


Figure 2 - ML process

Approaches for Demonstration: *Discussion of the capabilities and challenges associated with different ML technology that may affect adoption decisions for safety related RAS.*

2.3.1 Sufficiency of training

Assurance Objective: The learned algorithm is trained to satisfy the safety requirements using data that is sufficiently representative of the RAS operating environment and operating scenarios.

Contextual Description: The training data used must be sufficient to ensure that the trained algorithm will satisfy the defined safety requirements. This must include assurance that the training data provides sufficient coverage of all operating scenarios in the defined operating environment. At the same time, it must also be ensured that the machine-learned component does not become over-fitted to the training data resulting in lack of generalisation of the learning. This means that the machine-learned component should be shown to be robust, in that its performance with test data does not significantly deteriorate from the performance achieved with from the training data. Issues relating to the processing and classification of the training data are also important considerations.

Approaches for Demonstration*: The machine-learnt components may be trained through operation of the system itself, or may be trained on a simulator before integration into the target RAS. Machine-learnt components being used in safety related systems are often trained off-line prior to deployment and then may be updated during operation. The guidance will discuss the assurance considerations associated with real-world and simulation-based training (or the use of a combination).*

There are common challenges that may be encountered as part of the training process. As well as ensuring robustness, these include avoiding negative side effects, and avoiding 'reward hacking' and other potential problems when using a reinforcement based learning approach. It is important that the training undertaken can be demonstrated to mitigate such problems.

Explainability can be important as part of training by identifying what has been learned, and thus ways to make the training data more effective.

2.3.2 Sufficiency of the learning process

Assurance Objective: The learning approach is appropriate to satisfy the defined safety requirements.

Contextual Description: The approach taken to learning can affect assurance in a number of ways. The types of models, templates and parameters selected by the system developers during the machine learning process can all impact the satisfaction of the requirements. Decisions are also required on whether an off-line or on-line learning approach is most appropriate. There will also be uncertainty in the outputs of the learned algorithms (such as a stated confidence in a classification) that must be accounted for.

***Approaches for Demonstration**: TBD*

2.3.3 Verification of the learned model

Assurance Objective: Demonstrate that the learned model satisfies the defined safety requirements.

Contextual Description: It is necessary to generate evidence that provides sufficient confidence that the learned model will satisfy the relevant safety requirements throughout operation. This will require evidence regarding all defined operating scenarios in the defined operating environment. Evidence may be generated either through dynamic testing, or by static analysis of the learned algorithm.

***Approaches for Demonstration**: TBD*

For all testing approaches, the focus is on the sufficiency of the test data used with respect to coverage, and a requirement to be independent from the learning data. The machine-learnt components may be tested through operation of the system itself, or tested on a simulator before integration into the target RAS, or a mix of the two.

For all verification approaches there are challenges associated with the specification and verification of the assumptions it is necessary to make about the environment and operation in order to create usable models. Lack of explainability of learned models can make them hard to analyse.

2.4. Controlling interactions with other systems

Assurance Objective: Identify how interactions between the RAS and other systems may give rise to unsafe behaviour.

Assurance Objective: Manage interactions between the RAS and other systems to ensure they do not result in unsafe behaviour.

Contextual Description: Although a RAS in itself may considered to be safe, it may still be possible for unsafe behaviour to emerge due to the interactions between the RAS and other systems. This may be interactions with another RAS, or with 'manually' controlled systems. Such interacting and collaborating systems are often referred to as a "System of Systems (SoS)". For many applications, RAS will operate as part of a larger SoS (although RAS will not always be part of a SoS, and not all SoS will contain an RAS). It is required to provide sufficient confidence that potentially unsafe interactions between the RAS and other systems have been identified and mitigated.

Approaches for Demonstration: *TBD*

2.5 Controlling interactions at the System-level

Assurance Objective: Identify how interactions between elements of the RAS architecture may give rise to unsafe behaviour.

Assurance Objective: Manage interactions between elements of the RAS architecture to ensure they do not result in unsafe behaviour.

Contextual Description: Although the individual between elements of the RAS architecture may meet their requirements, it may still be possible for unsafe behaviour to emerge due to the interactions between those elements. It is therefore required to provide sufficient confidence that potentially unsafe interactions between elements have been identified and mitigated.

Approaches for Demonstration:

As part of addressing these objectives, the impact of limitations of one element of the architecture on the behaviour of another must be considered, for example design decisions taken on how to implement sensing may impact on the performance that is achieved by functions such as object classification.

2.6 Handling change during operation

Assurance Objective: Ensure that the RAS responds safely to changes that occur during operation.

Contextual Description: In comparison to more traditional systems, it is expected that there will be more changes during the operation of an RAS that are unpredictable during the development of that system. Example operational changes for RAS include changes to the behaviour of the system due to adaptation as a result of learning.

Approaches for Demonstration:

2.6.1 Monitoring RAS operation

Assurance Objective: Identify changes that occur during the operation of the RAS that may result in unsafe behaviour.

Contextual Description: Mechanisms are required to be in place to monitor for potentially unsafe changes. It is important to identify what must be monitored during system operation in order to assure its continued safe operation. This will often be identified from considering assumptions and context defined as part of the assurance case for the RAS. Where it is identified that the assumptions or context must hold in order for the system to be considered safe, and they may become invalid during operation (e.g. certain visual sensors may require minimum lighting levels), then those need to be monitored.

Approaches for Demonstration: *TBD*

2.6.2 Defining safe system response to changes

Assurance Objective: Define the safe response required of the RAS when potentially unsafe changes are identified.

Contextual Description: Once potentially unsafe changes are detected, a safe response must be enacted (i.e. returning the system to a safe state). What is an appropriate response will depend upon the nature of the change that occurs and must link back to the higher-level safety analysis of the RAS. For example, for some changes it may be determined that the safest response is to hand back control to an operator; for other changes this may be an unsafe response.

Approaches for Demonstration: *TBD*

2.7 Using Simulation

Assurance Objective: Ensure that simulations used as part of the assurance process provide a representation of the real world sufficient for their use.

Contextual Description: Simulation may be used in a number of different roles as part of assuring RAS (such as training of ML systems, testing and understanding system behaviour). In all cases it is important for assurance that a sufficient level of correspondence can be demonstrated between the simulation model and the real-world that it models. What is sufficient will depend on what is being modelled and why. For example simulations of the sensing functions may require a detailed correspondence in the simulation to raw sensor data from the real sensors as well as an accurate model of environmental effects, whereas a detailed model of the vehicle itself and its dynamic behaviour may not be required. Simulation may include hardware-in-the-loop approaches where simulated inputs are provided to real physical systems.

***Approaches for Demonstration**:*

General guidance on how to create good simulations and judge their sufficiency, including the creation of 'digital twins' through the use of real-world data to create simulation models.

Specific guidance on the use of simulation in specific roles will be provided against the particular assurance objective.

2.8 Explainability

Assurance Objective: Be able to provide explanations, when required, for decisions taken by the system.

Contextual Description: It is often important for assurance that explanations can be provided as to why a particular decision was taken by the system in a particular set of circumstances. There are four main reasons why explainability is important:

- Explain to justify – It may be necessary as part of the assurance or regulatory process to provide a justification for why a particular decision was taken.
- Explain to correct – When an algorithm is being trained, in order to improve its performance, it may be necessary to correct errors that are made by the algorithm (such as mis-classification). Correcting errors successfully may require explanation of why the incorrect decision was made by the algorithm.
- Explain to improve – If the performance of an algorithm needs to be improved, an explanation of decisions taken may help to identify how improvements can be achieved most effectively.

- Explain to discover – To ensure that the learning process is effective, it may be necessary to have an understanding of parameters or characteristics that have a significant impact on what is learned.

***Approaches for Demonstration**:*

General guidance on how to ensure decisions are explainable in a manner that is comprehensible by a human.

Specific guidance on the use of explainability for specific goals will be provided against the particular assurance objective.

3 Understanding and controlling deviations from required behaviour

Assurance Objective: Deviations from required behaviour during operation will not result in unacceptable safety risk.

Contextual Description: Even if sufficient effort is made to implement a system that satisfies all the safety requirements, it is still necessary to also explicitly consider the ways in which the system may deviate from that required behaviour during operation. Deviations may arise due to random failures (such as component degradation during operation) or systematic failures (such as design errors) in the system. They may also arise as a result of security attacks on the system. To provide assurance, the potential for unsafe deviations must be identified and mitigated as considered in the following sub-objectives described below.

***Approaches for Demonstration**: TBD*

3.1 Identifying potential deviation from required behaviour

Assurance Objective: Identify potential sources of deviation from required behaviour.

Contextual Description: Deviations may occur in any of the element of the RAS architecture (Sensing, Understanding, Deciding, Acting and Infrastructure). The potential deviations, and their impact on the satisfaction of safety requirements must be identified for each of the elements and the RAS as a whole.

***Approaches for Demonstration**:*

There are standard methods of identifying deviations (such as HAZOP) which may be applied to meet this objective. The way in which this objective is demonstrated may be different for different elements of the architecture. There may, for example be common failure modes associated with particular technologies that must be managed for each element, as defined in the following sub-sections.

3.1.1 Identifying 'Sensing' deviations

***Approaches for Demonstration**:*

Guidance on common failure modes for different types of sensor.

3.1.2 Identifying 'Understanding' deviations

***Approaches for Demonstration**: TBD*

3.1.3 Identifying 'Deciding' deviations

***Approaches for Demonstration**: TBD*

3.1.4 Identifying 'Acting' deviations

***Approaches for Demonstration**:*

Guidance on common failure modes for different types of actuator.

3.1.5 Identifying Infrastructure deviations

***Approaches for Demonstration**: TBD*

3.1.6 Identifying ML deviations

Assurance Objective: Identify potential sources of deviation from required behaviour for the machine-learnt components of the system.

Contextual Description: Although sufficient effort is made to provide a machine-learnt component that satisfies all the safety requirements, it is still necessary for assurance to also explicitly consider mechanisms that might cause a machine-learnt component to deviate from that implementation during operation. This may include, for example, mechanisms resulting in false positive or false negative classifications as part of the understanding function. In comparison to more traditional systems, it is often more challenging to identify deviations in machine-learnt components (due to issues of explainabilty).

***Approaches for Demonstration**: TBD*

3.1.7 Interaction deviations

Assurance Objective: Identify potential sources of deviation from required behaviour resulting from the interactions between elements of the system.

Contextual Description: Deviations may occur as a result of the interactions between system elements. These deviations would not be identified through considering each element in isolation. These may arise, for example, as a result of inconsistent assumptions regarding different elements of the system.

***Approaches for Demonstration**: TBD*

3.1.8 Human/Machine interactions

Assurance Objective: Identify potential sources of deviation from required behaviour resulting from the interactions between humans and the system.

Contextual Description: Deviations may occur as a result of the interactions between humans (whether the operator or another human in the environment of the system) and the system itself. Many functions of an RAS may be implemented using a combination of machines and humans. For example, some decisions may be allocated to software, whilst some may remain with a human. The allocation of decisions may also change during system operation depending upon particular scenarios or the state of the system. The potential deviations caused by these interactions, and their impact on the satisfaction of safety requirements must be identified.

***Approaches for Demonstration**: TBD*

3.2 Mitigating potential deviations

Assurance Objective: Manage potential sources of deviation to ensure they do not result in unsafe behaviour.

Contextual Description: For identified sources of deviation, a sufficient mitigation must be identified and implemented. What is an appropriate mitigation will depend upon the nature of the deviation, and must link back to the higher-level safety analysis of the RAS. Mitigation may require additional safety requirements to be derived and implemented.

Approaches for Demonstration:

There are generic strategies for managing failures such as:

- *Human as a back-up (Handover)*
- *Redundancy and Diversity*
- *Degraded modes of operation*
- *Reversion to a defined safe state*

Guidance must consider the appropriateness and potential limitations of such strategies when applied to RAS.

Where appropriate, specific mitigation strategies for particular types failure modes may also be adopted.

Where mitigations are identified, corresponding requirements must be defined and implemented to assure that the mitigations are put in place.

3.2.1 Managing failures of machine-learnt components

Assurance Objective: Implement effective mitigations for identified sources of deviation for machine-learnt components of the system.

Contextual Description: Any identified mechanism that could cause a machine-learnt component to deviate from its intended behaviour during operation must be shown to be effectively managed. This may necessitate the definition of additional requirements for implementation, or changes to the system design.

***Approaches for Demonstration**: TBD*

3.2.2 Managing assurance deficits

Assurance Objective: Manage assurance deficits to ensure they will not present an unacceptable risk.

Contextual Description: Systematic failures may occur where there are gaps in the information or knowledge about the system and its behaviour (epistemic uncertainty). These gaps can be referred to as assurance deficits. Although there will always be assurance deficits (as it is not possible to have complete knowledge of the system and its environment), it is important that where there are known to be assurance deficits, they are sufficiently managed to ensure they do not present an unacceptable safety risk to the system.

Approaches for Demonstration:

Guidance on generic strategies for managing assurance deficits such as:

- *Provide mitigation through system design or requirements change*

- *Provide mitigation through operational constraints or restrictions*
- *Generate additional information to "fill the knowledge gap"*
- *Accept the risk associated with the assurance deficit*

With reference to the second bullet point above, it may be that additional restrictions are put in place to ensure that the risk is acceptable given the known uncertainty. As more is learnt about the system through operation, it may then be possible to reduce or remove such restrictions, as the associated risk has reduced. This incremental approach enables assurance to be increased through operation, without exposure to intolerable risk. Guidance will be provided on incremental assurance.

4 Gaining approval for operation of RAS[2]

Regulatory Objective: Identify the entity or entities from which approval is required prior to operation of the RAS.

Regulatory Objective: Gain approval for the operation of the RAS in the defined operational and environmental context.

Contextual Description: There may be a number of different entities or stakeholders from which approval for operation may be required. For domains with an explicit and established regulatory regime, the approving entity may be easily identified, in other domains this may not be the case.

The approval of an RAS may be subject to certain provisions or restrictions on operation. This objective is achieved through the consideration of the sub-objectives as described below.

Approaches for Demonstration:

4.1 Conforming to rules and regulations

Regulatory Objective: Demonstrate that all the applicable requirements of the rules and regulations have been conformed to.

Contextual Description: Evidence must be generated to demonstrate that all the applicable requirements have been conformed to. Where interpretation of the requirements has been required, or where an alternative means of compliance has been adopted, a reasoned compliance justification may be required. This objective is achieved through the consideration of the sub-objectives as described below.

[2] Regulatory aspects will be expected to contain a higher proportion of domain specific information. At this stage, the material here is generic and domain independent.

Approaches for Demonstration: *TBD*

4.1.1 Identifying applicable rules and regulations

Regulatory Objective: Correctly identify all of the rules and regulations that are applicable to the operation of the RAS in its defined operating environment and scenarios.

Contextual Description: The applicable rules and regulations that an RAS must conform to will vary depending upon the context of its use, or the environment in which it is operating.

Approaches for Demonstration: *TBD*

4.1.2 Understanding the requirements of rules and regulations

Regulatory Objective: Correctly interpret the requirements of the applicable rules and regulations.

Contextual Description: The applicable rules and regulations will in most cases have been created on the assumption of a more traditional system. The implications of the rules and regulations in the context of an RAS must be correctly determined. In some cases it may be determined that a particular requirement of an applicable regulation cannot be met for an RAS. In such cases an acceptable alternative means of compliance may need to be determined and agreed.

Approaches for Demonstration: *TBD*

4.2 Risk Acceptance

Regulatory Objective: Gain acceptance for the residual safety risk associated with the RAS operation.

Contextual Description: Safety engineering activities will mitigate the risks associated with the operation of the RAS. Approving operation will require that the level of residual risk remaining following mitigation is accepted by the approving entities.

This objective requires the consideration of the further sub-objectives described below.

Approaches for Demonstration:

Some regulatory regimes have very clear criteria for judging the acceptability of risk reduction (such as ALARP). In other cases a more ad-hoc justification for the acceptability of risk may be required.

4.2.1. Evaluating risks and benefits of RAS operation

Regulatory Objective: Justify trade-offs that are made of risks and benefits associated with the operation of the RAS.

Contextual Description: The utilisation of autonomy for previously manual tasks has the potential to introduce new risks, however autonomy also has the potential to reduce the risks that were previously associated with the manual task (for example by removing the exposure of an operator to a hazard, or enabling more reliable performance of a complex task). When making a decision to approve the operation of the RAS, a risk-benefit trade-off is therefore often required. Although risk-benefit assessment may sometimes be undertaken for more traditional systems (particularly in certain domains such as medical), it is more likely that such an assessment will be required for an RAS.

As part of making a justifiable decision, the uncertainty in the assessment of risk and benefit must be considered.

Approaches for Demonstration: *TBD*

4.2.2. Consideration of ethical issues

Regulatory Objective: Demonstrate that all relevant ethical issues associated with the operation of the RAS have been adequately addressed.

Contextual Description: The operation of RAS will often bring into consideration ethical issues that are not present for traditional systems, mainly resulting from a need to encode previously implicit decision making processes. These ethical considerations, which must include possible biases introduced through ML, must be addressed in a way that is acceptable to the approving entities.

Approaches for Demonstration: *TBD*

4.3 Provision of sufficient confidence in the required behaviour

Regulatory Objective: Demonstrate that there is sufficient confidence that the system's behaviour will be sufficiently safe throughout its life.

Contextual Description: The assurance objectives (areas 1 through 3) set out how the assurance of an RAS may be demonstrated. The level of confidence with which the objectives are demonstrated will vary depending upon the assurance activities undertaken and the nature and amount of evidence that is provided. The confidence provided must be sufficient to enable a defensible decision to be made on allowing the operation of the RAS.

The judgement on what level of confidence is required against the objectives is subjective. For some domains, guidance on making such judgements may be available through standards and other guidance documents. For other domains judgements may need to made on a case-by-case basis using a risk-based approach (as considered in 4.2).

Approaches for Demonstration: *TBD*

4.4 Provision for investigation of incidents and accidents

Regulatory Objective: Provide means to support the investigation of incidents and accidents that may occur when operating the RAS.

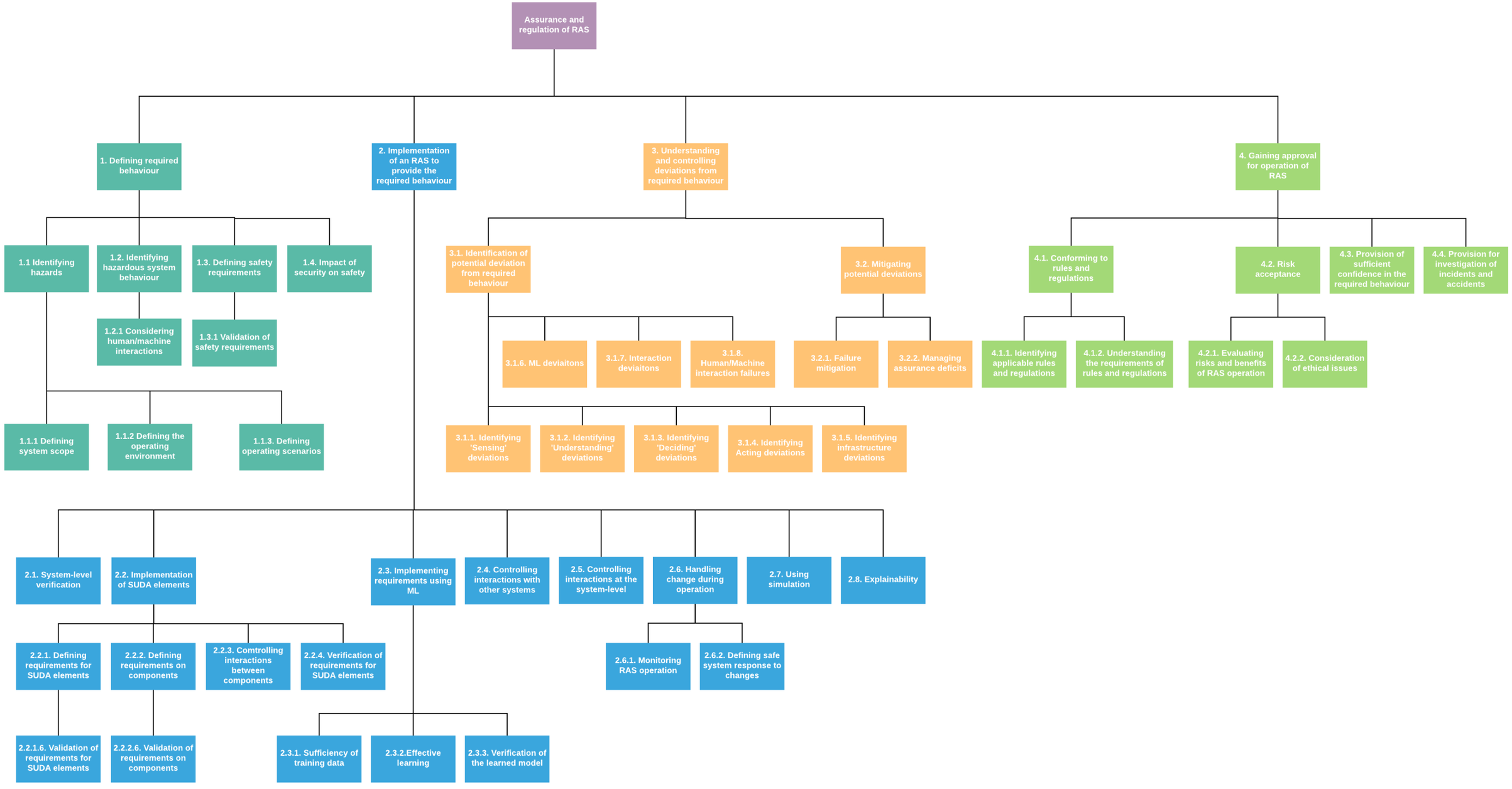
Contextual Description: It is important to be able to learn from incidents and accidents in order that they can be prevented from happening again (the risk reduced). This may require that measures are put in place to facilitate this investigation. For RAS, this will often require that investigators (a) have access to data generated by the RAS prior to the accident and (b) are able to interpret that data to determine the causes of the accident or incident. This may result in additional requirements on the RAS to implement a suitable information collection mechanism.

As a result of such investigations it may be required to update already deployed RAS with the knowledge that has been gained (e.g. such that the RAS is able to predict or detect similar situations). This will require that RAS are designed and implemented so they can be updated.

Approaches for Demonstration: *TBD*

3. Overview of BoK Structure

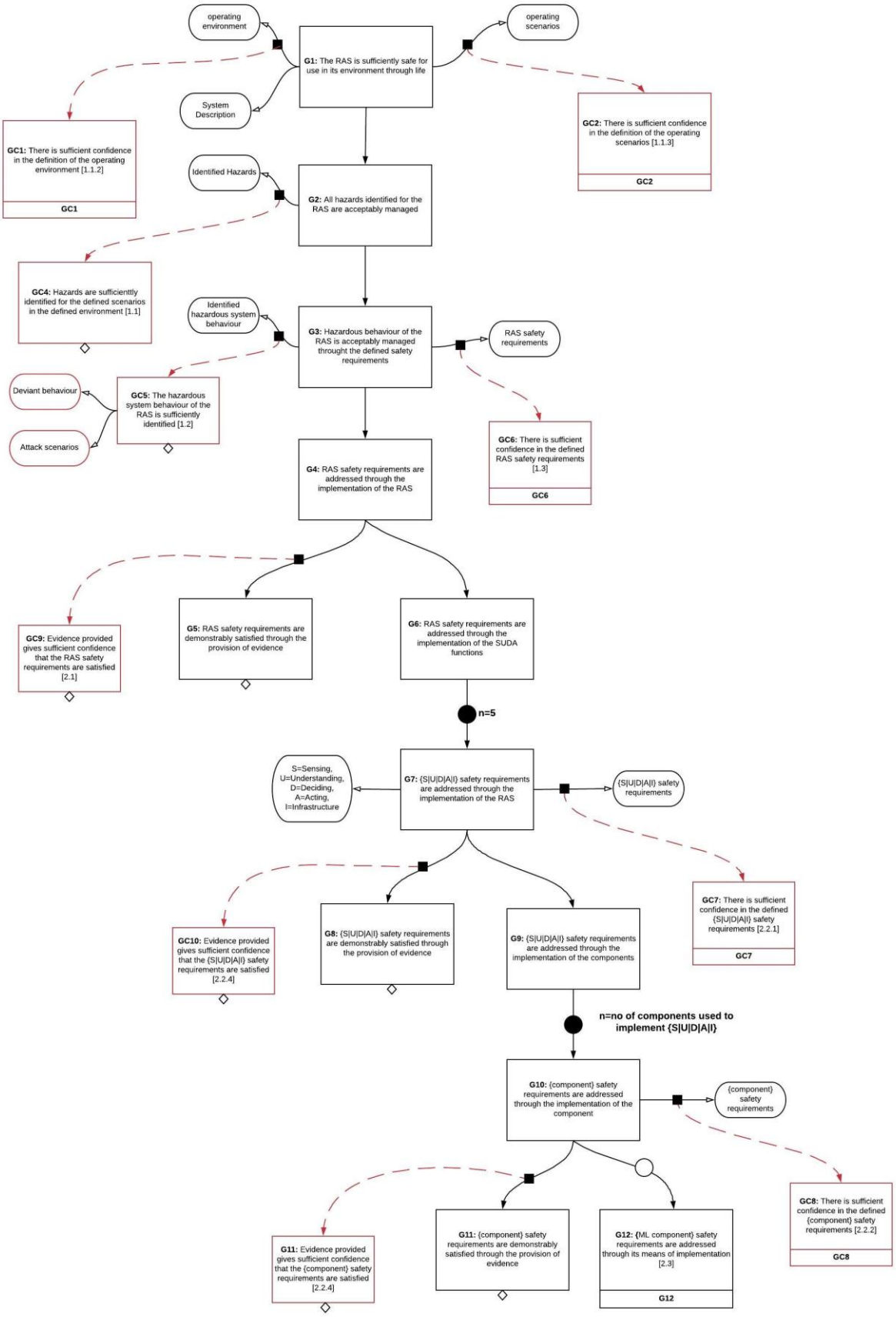
The diagram below illustrates the structure of the BoK as described above. Such a structure could be used as a graphical interface to help users locate the relevant information with the BoK.

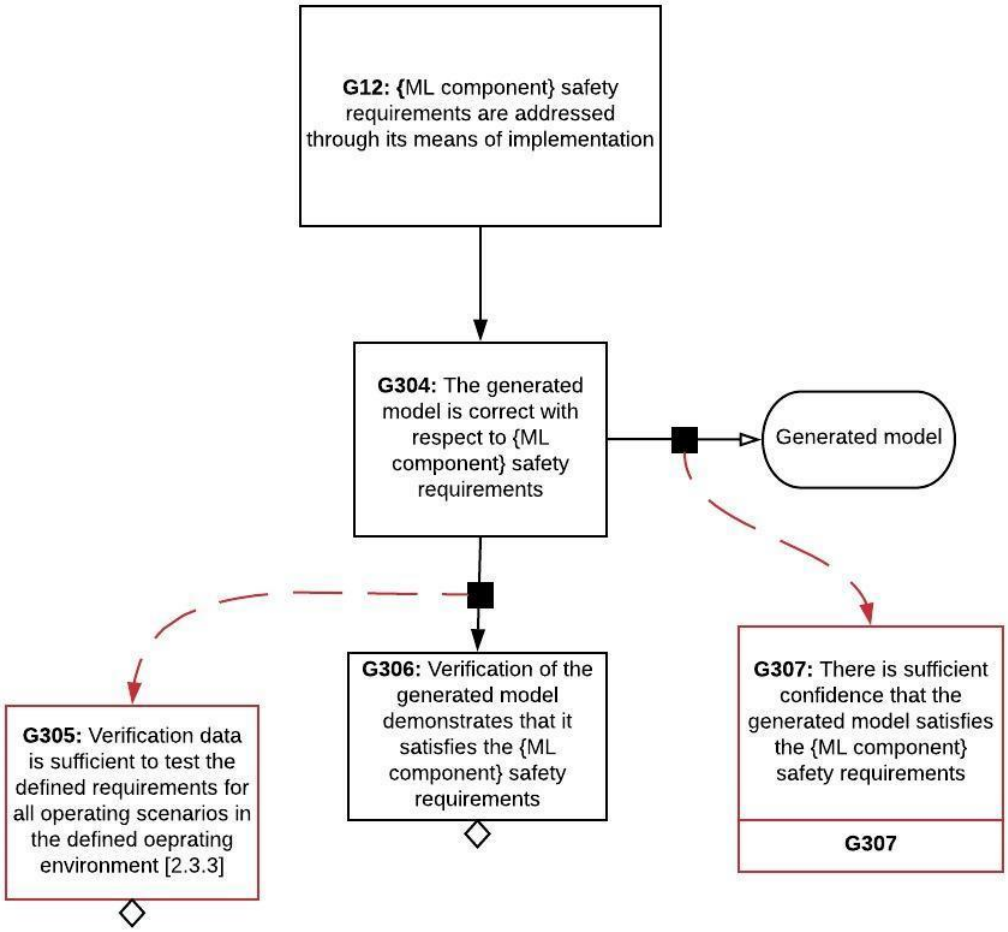


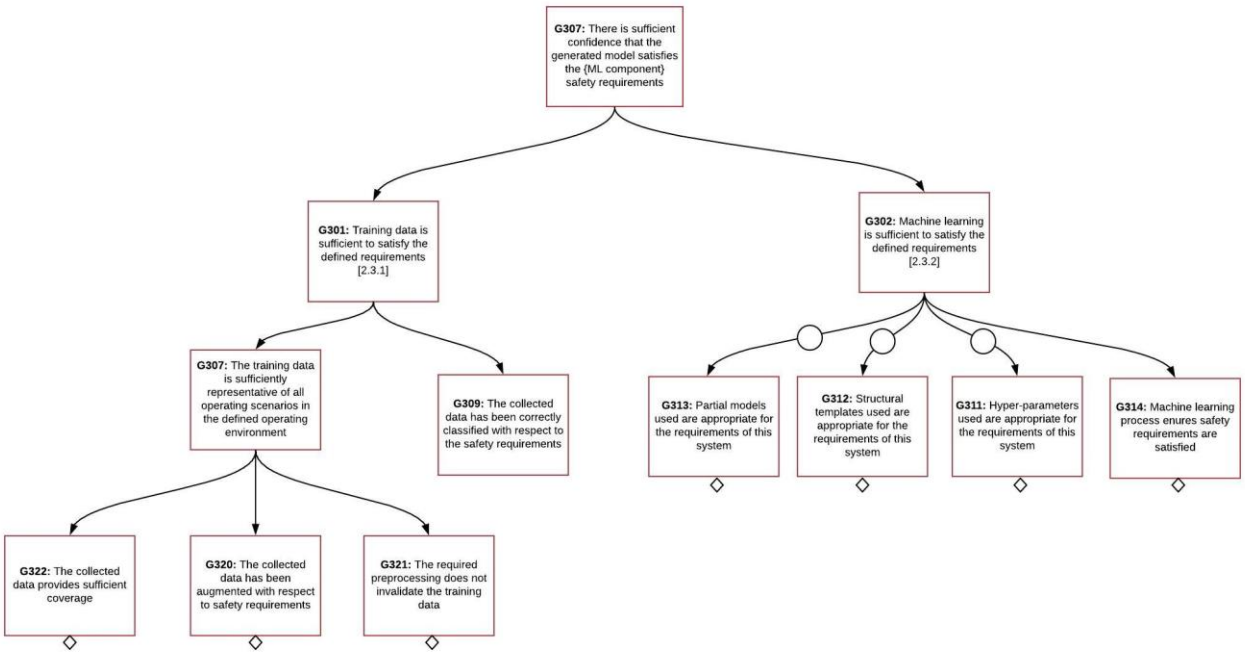
4. Assurance case for RAS

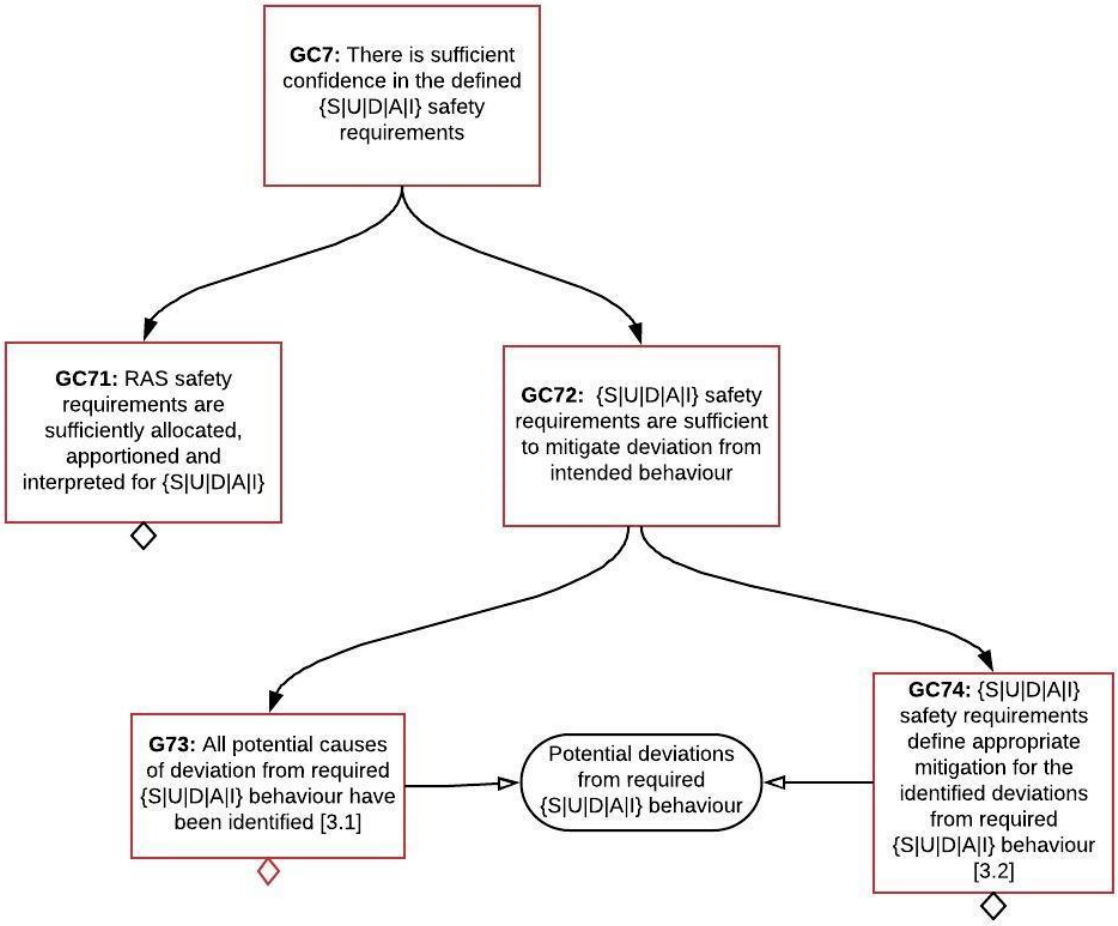
Assurance cases are often used to demonstrate the safety of systems. Ultimately, the guidance that will be provided in the BoK using the structure described above should support the development and evaluation of a compelling assurance case for the through-life operation of the RAS. This section illustrates how this can be achieved by provided a suggested structure for the assurance argument in the form of an assurance case pattern represented using Goal Structuring Notation (GSN)[3]. For each of the required elements of the assurance case pattern, corresponding guidance will be available in the BoK. This is indicated by the section references provided in argument claims in the figures below. The intention is that an interested party could click on the reference within the pattern structure to locate the BoK guidance relating to supporting that safety claim.

[3] SCSC Assurance Case Working Group, GSN Community Standard. Version 2. January 2018.


operating environment
G1: The RAS is sufficiently safe for use in its environment through life
operating scenarios
System Description
GC1: There is sufficient confidence in the definition of the operating environment [1.1.2]
GC1
GC2: There is sufficient confidence in the definition of the operating scenarios [1.1.3]
GC2
Identified Hazards
G2: All hazards identified for the RAS are acceptably managed
GC4: Hazards are sufficienttly identified for the defined scenarios in the defined environment [1.1]
Identified hazardous system behaviour
G3: Hazardous behaviour of the RAS is acceptably managed throught the defined safety requirements
RAS safety requirements
Deviant behaviour
GC5: The hazardous system behaviour of the RAS is sufficiently identified [1.2]
Attack scenarios
GC6: There is sufficient confidence in the defined RAS safety requirements [1.3]
GC6
G4: RAS safety requirements are addressed through the implementation of the RAS
GC9: Evidence provided gives sufficient confidence that the RAS safety requirements are satisfied [2.1]
G5: RAS safety requirements are demonstrably satisfied through the provision of evidence
G6: RAS safety requirements are addressed through the implementation of the SUDA functions
n=5
S=Sensing, U=Understanding, D=Deciding, A=Acting, I=Infrastructure
G7: {S|U|D|A|I} safety requirements are addressed through the implementation of the RAS
{S|U|D|A|I} safety requirements
GC10: Evidence provided gives sufficient confidence that the {S|U|D|A|I} safety requirements are satisfied [2.2.4]
G8: {S|U|D|A|I} safety requirements are demonstrably satisfied through the provision of evidence
G9: {S|U|D|A|I} safety requirements are addressed through the implementation of the components
GC7: There is sufficient confidence in the defined {S|U|D|A|I} safety requirements [2.2.1]
GC7
n=no of components used to implement {S|U|D|A|I}
G10: {component} safety requirements are addressed through the implementation of the component
{component} safety requirements
G11: Evidence provided gives sufficient confidence that the {component} safety requirements are satisfied [2.2.4]
G11: {component} safety requirements are demonstrably satisfied through the provision of evidence
G12: {ML component} safety requirements are addressed through its means of implementation [2.3]
G12
GC8: There is sufficient confidence in the defined {component} safety requirements [2.2.2]
GC8


G12: {ML component} safety requirements are addressed through its means of implementation
G304: The generated model is correct with respect to {ML component} safety requirements
Generated model
G305: Verification data is sufficient to test the defined requirements for all operating scenarios in the defined oeprating environment [2.3.3]
G306: Verification of the generated model demonstrates that it satisfies the {ML component} safety requirements
G307: There is sufficient confidence that the generated model satisfies the {ML component} safety requirements
G307


G307: There is sufficient confidence that the generated model satisfies the {ML component} safety requirements
G301: Training data is sufficient to satisfy the defined requirements [2.3.1]
G302: Machine learning is sufficient to satisfy the defined requirements [2.3.2]
G307: The training data is sufficiently representative of all operating scenarios in the defined operating environment
G309: The collected data has been correctly classified with respect to the safety requirements
G313: Partial models used are appropriate for the requirements of this system
G312: Structural templates used are appropriate for the requirements of this system
G311: Hyper-parameters used are appropriate for the requirements of this system
G314: Machine learning process enures safety requirements are satisfied
G322: The collected data provides sufficient coverage
G320: The collected data has been augmented with respect to safety requirements
G321: The required preprocessing does not invalidate the training data


GC7: There is sufficient confidence in the defined {S|U|D|A|I} safety requirements
GC71: RAS safety requirements are sufficiently allocated, apportioned and interpreted for {S|U|D|A|I}
GC72: {S|U|D|A|I} safety requirements are sufficient to mitigate deviation from intended behaviour
G73: All potential causes of deviation from required {S|U|D|A|I} behaviour have been identified [3.1]
Potential deviations from required {S|U|D|A|I} behaviour
GC74: {S|U|D|A|I} safety requirements define appropriate mitigation for the identified deviations from required {S|U|D|A|I} behaviour [3.2]

5. Definitions

In this section we define terms used in the assurance objectives in the section 2. Where alternative definitions are required as part of guidance material in the BoK (for example if domain specific guidance uses the term 'Hazard' in a different way) these terms may be redefined for that purpose, but the standard definitions below should remain stable as the default throughout the BoK.

Assurance (n) - Justified confidence in a property.

Safety assurance - Justified confidence in safety.

Safety - The degree of freedom from hazard risk.

Hazard - A condition of a system that can develop into an accident through a sequence of normal events and actions.

Accident - An unintended event or sequence of events leading to harm.

Risk - The product of severity and probability.

Hazard risk - The product of the severity and probability of a hazard.

System - A group of interacting or interrelated elements that form a unified whole.

Component - Element that forms part of a system.

Regulation (n) - A set of rules or directives.

Regulatory authority - An organisation that can make, maintain or enforce regulations.

Autonomy - The capability to make decisions free from human control.[4]

Autonomous - Having autonomy.[5]

Robotics - The design, construction, operation, and use of robots.

Robot - A machine capable of carrying out a complex series of actions automatically.

Automatic - Able to operate independently of human control.[6]

Machine Learning (ML) - A process by which computers create a model of the real-world from data in the form of observations and real-world interactions.

Assurance case - Arguments and evidence intended to demonstrate assurance.

Verification - The evaluation of compliance to a specification.

[4] See further discussion in section 5.1 below of the programme's use of the term 'Autonomy'.
[5] See further discussion in section 5.1 below of the programme's use of the term 'Autonomous'.
[6] See further discussion in section 5.1 below of the programme's use of the term 'Automatic'.

Validation - The evaluation of the correctness of a specification.

Hazardous behaviour - Behaviour that may result in a hazard.

Attack scenario - An event or sequence of events through which a vulnerability may be exploited.

Vulnerability - A weakness which can be exploited to perform an attack against assets.

Safety requirement - Description of a property or behaviour required to ensure safety.

Security threat – An intentional or unintentional event that may exploit a vulnerability in a system and result in harm.

Vulnerability - A weakness in a system that can be exploited by an attacker to perform unauthorized actions on that system.

Safety justification - An evidence-based justification of safety assurance.

Formal verification - Verification using mathematical methods.

Reinforcement learning - A type of machine learning that allows computers to determine their required behaviour through exploration within a specific context, in order to maximise some notion of cumulative reward.

Testing - Evaluation through operation.

Static Analysis - Evaluation without operation.

Simulation - A model of a real-world situation on a computer.

Random failure - Failure due to random events, most commonly resulting from physical causes, that can be characterised by statistical failure models.

Systematic failure - Failure due to flaws in specification, design, manufacture, installation or maintenance.

Failure mode - A specific way in which failure may occur.

Assurance deficit - A specific source of epistemic uncertainty caused by a lack of knowledge or information.

Conformance - Fulfillment of requirements.

Residual risk - The risk that remains once all risk reduction measures have been taken.

Incident - An event which significantly degrades safety margins, but does not lead to an accident.

Assurance argument - An argument used to demonstrate assurance based upon the available evidence.

Argument - A series of claims intended to establish the truth of a conclusion.

Assurance case pattern - A means of documenting and reusing assurance argument structures.

5.1 Further Discussion of 'Autonomy'

The Programme takes the view that the key difference between manually controlled and autonomous systems is that the RAS has decision-making capability and authority. This is what is meant by decisions free from human control. All software implements decisions in a sense, e.g. taking an *else* rather than a *then* branch. However, the intent is that the decisions are those that might otherwise have been taken by humans and that require intelligence, situational understanding and freedom, in the sense of individual autonomy, e.g. stopping at a red light, or categorising an object as a person rather than a lamp-post.

The notion of "taken by humans" is not sharply defined, and we might define some systems, e.g. a kettle which shuts-off when the water is boiling, as automatic not autonomous. In general, we would expect the term autonomy, rather than automatic, to be used where:

- there is an open environment, e.g. as in driving on the roads, as opposed to a closed environment which is well-defined and understood;
- the range of options in decision-making is very large and may not even be bounded;
- there is considerable uncertainty in assessing the situation and/or choosing a course of action (making a decision).

In practice, the BoK will provide guidance in a way which reflects the particular challenges, e.g. open vs closed environments, and will not be constrained by whether or not some RAS is viewed as automatic as opposed to autonomous.

In many domains standards or other documents define levels of autonomy from full human control, via shared human-machine decision-making (or the possibility of handover from machine to human), up to "full autonomy", consistent with the definition given above. The intent is that the definition is interpreted flexibly, and would include shared human-RAS decision-making, not just "full autonomy".

Dictionary definitions of autonomy use phrases like "freedom from influence and control". We have deliberately excluded "influence" as we would expect RAS to be influenced by the operating environment, e.g. behaviours of other cars or pedestrians in autonomous driving, and behaviour of other ships in maritime autonomy.