Automated Vehicle Safety Assurance: A Framework for Automated Driving Systems

Abridged version of a report commissioned by the Department for Transport

Date: 15/05/2023

© HORIBA MIRA Ltd 2023. All rights reserved, subject to client contract. Information contained in this document may not be published in any form of advertising or other matter without prior agreement of the Chief Executive Officer of HORIBA MIRA.

HORIBA MIRA Ltd. Registered Office: Watling Street · Nuneaton · Warwickshire · CV10 0TU · England · www.horiba-mira.com
Tel: +44 (0)24 7635 5000 · Fax: +44 (0)24 7635 8000 Registered in England No 9626352 · VAT Registration GB 100 1464 84

Disclaimer

The information contained within this response does not necessarily represent the position of the Department for Transport.

Introduction

This report summarises recommendations for a safety and security assurance framework to regulate low-speed automated vehicles (LSAVs). It has been prepared by HORIBA MIRA (lead partner), University of York, TRL and Five as part of a project commissioned by the UK Department for Transport (DfT); this is an abridged version of the full report provided to DfT. Whilst initially aimed at LSAVs, it is anticipated that the findings could be extrapolated to other automated vehicle use cases, such as those at higher operating speeds.

A prescriptive approach to LSAV approval has *not* been recommended due to the rapidly evolving state of the art and the lack of a standardised safety assurance method. Instead, the submission of safety case reports to the regulator is proposed; this would permit the flexibility to employ alternative safety assurance solutions, whilst still providing an appropriate and comprehensive safety record to enable robust scrutiny.

A safety case report would comprise of: evidence to define the nature of the vehicles and their operation; evidence of appropriate safety analysis; evidence of appropriate testing and evaluation (verification and validation); and evidence of appropriate safety management systems. It must also contain a ‘safety argument’: a structured description of how the evidence is sufficiently complete and comprehensive such that, when all articles are considered together, they support the claim that the overall safety of the LSAV type is acceptable.

In this report, we assumed a commercial model that consists of a Manufacturer of the LSAV and an Operator of the mobility service. The proposed process would require the Manufacturer to submit a **system** safety case report and the Operator to submit a **deployment** safety case report, although it is permissible for the Manufacturer and the Operator to be the same organisation; indeed, it is anticipated that this may be the dominant model for early commercial deployments of LSAVs. The safety case reports would provide the regulator with all the necessary information, without providing the full information contained within the safety cases as developed and maintained by the Manufacturer and the Operator, which may be impractical to scrutinise.

The complete regulatory lifecycle consists of the phases:

- **Pre-Approval** – the engineering activities to develop the system and to acquire safety evidence prior to the application for approval;
- **Vehicle Type Approval** – the formal process of assessing the safety of the automated vehicle;
- **Deployment Approval** – the formal process of assessing the safety of a vehicle’s operation in its intended deployment environment;
- **Monitoring** – capturing data while the vehicles are in service to validate safety case assumptions and to identify where remedial action is required;
- **Response** – the implementation of remedial actions;
- **Change** – proposals to adapt or improve the vehicle capability or the service.

Definition of the System and Deployment

As a precursor to the safety evidence that is collated downstream, the nature of the system and its operation must be robustly defined. In addition to an ODD (Operational Design Domain), this report proposes that a ‘TOD’ (Target Operating Domain) must also be defined: while the former represents the *design intent*, the latter represents the *deployment reality*; this distinction is important since the two may not be identical.

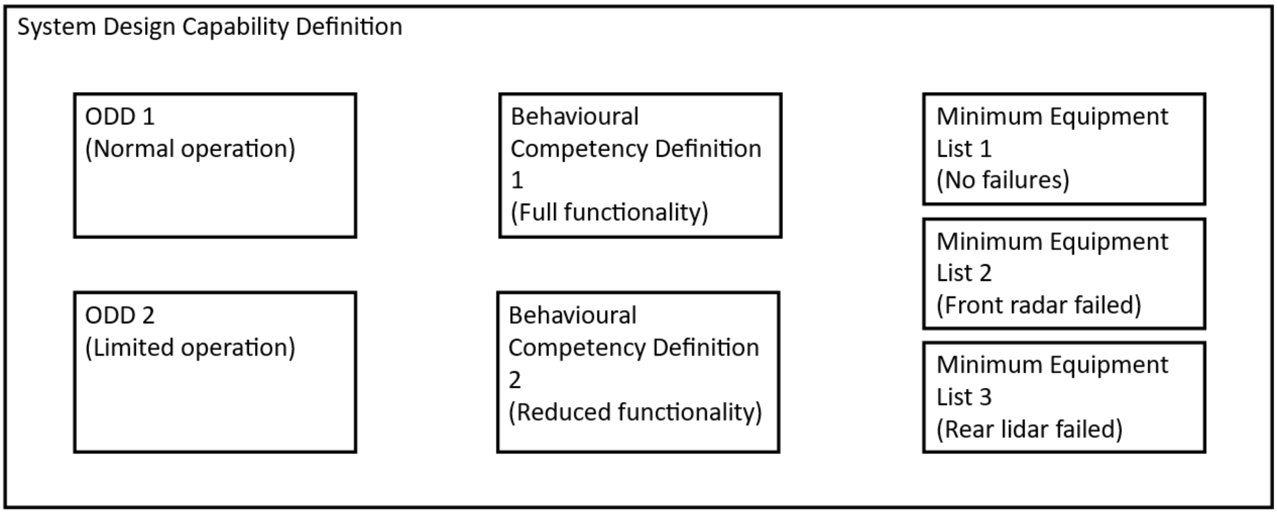
It is proposed that the TOD should include a definition of the specific location of the deployment route(s) or geofenced area(s), such that the actual location of the deployment within the world is unambiguously defined, rather than merely described in an abstract manner by generic attributes that could apply to many locations. The ODD may optionally be specific to the actual deployment route(s), or it may be generic such that a system developed to operate within it is compatible with multiple specific TODs. However, care should be taken to ensure early in the process that there are viable real-world deployment locations with a TOD that is compatible with the ODD the vehicle is being developed and assured for, to avoid the risk of investing in a vehicle that has no practicable applications.

The reason for requiring that a TOD be specific to a defined location is that some elements of the safety case are specifically linked to the deployment location. For example, a review of the operational safety of the route, such as identifying particular segments which may pose a hazard, or a review of the impact on traffic flows, would both be specific to the deployment location. Furthermore, a significant proportion of the testing of the full system should be conducted upon the actual deployment route (and potentially upon a representative equivalent such as a 'digital twin' within a simulation or an accurate mock-up within a proving ground). This is because even subtle differences between the actual location and a generic road layout, such as a piece of infrastructure that affects line of sight or a different junction geometry, could have a significant effect upon the LSAV's behaviour.

It would not, for example, be permissible to test a system using solely locations within Greenwich and Coventry, and then approve the system as safe for deployment on a route in Milton Keynes upon which it has never been tested; the range of road permutations that exists in the world, and the challenge of identifying and testing the system's response to them, are too great for 'go-anywhere' approvals to be practicable within the foreseeable future. This is in line with the existing state of the art, such as pilot deployments of driverless vehicles in the USA that are tested extensively within specified areas or routes prior to driverless operation taking place within those specific locations only.

The definition of the system must include a set of 'Behavioural Competencies' defining the functionalities that the system is required to perform, and a 'Minimum Equipment List' (MEL) of subsystems that must be in a fault-free state for the system to operate correctly in a given mode. It is anticipated that a system may have multiple definitions of ODD/TOD, Behavioural Competencies and MELs that can be combined in different ways to facilitate operation in degraded modes (e.g. to accommodate faults or adverse weather by restricting speed, eliminating certain manoeuvres or avoiding some parts of the route). The permissible combinations of ODD/TOD, Behavioural Competencies and MELs should be documented, e.g. within a matrix format. An example of such an approach is shown in Figure 1.

All foreseeable situations that the ADS would need to detect *and* take immediate action to must be defined as inside the ODD and TOD that the system is operating within at any given time, in order to ensure that such permutations are given appropriate attention within the vehicle design, analysis and testing. For example, operation in heavy snow could be defined as outside the ODD or TOD because, while the system would need to be able to detect it, there would be no need to take emergency action. On the other hand, if the presence of an e-scooter and rider cannot be ruled out for a deployment on a university campus, they should be defined as within the ODD and TOD, even if their presence results in degraded performance (e.g. reducing speed or having a wider zone around the vehicle in which objects will trigger a stop), to ensure that the safe response to e-scooter scenarios is properly analysed and tested. The ODD and TOD must be validated as shown in Figure 2.


System Design Capability Definition
ODD 1 (Normal operation)
ODD 2 (Limited operation)
Behavioural Competency Definition 1 (Full functionality)
Behavioural Competency Definition 2 (Reduced functionality)
Minimum Equipment List 1 (No failures)
Minimum Equipment List 2 (Front radar failed)
Minimum Equipment List 3 (Rear lidar failed)

Permutation Number	Permutation Name	ODD Definition	Behavioural Competency Definition	Minimum Equipment List	Permitted Combination?	Justification
1	Ideal Conditions	1	1	1	Yes	Normal, intended operation
2	-	1	1	2	No	Full behaviour not possible with failure
3	-	1	1	3	No	Full behaviour not possible with failure
4	-	1	2	1	No	Not necessary to restrict behaviour – no failures
5	-	1	2	2	No	Not possible to proceed with radar sensor unless ODD restricted
6	No Reversing	1	2	3	Yes	Reduced functionality to allow for failed sensor
7	-	2	1	1	No	Limited ODD unnecessary with no failure
8	Radar Failed	2	1	2	Yes	ODD restriction sufficient to allow full behaviour with radar fault
9	-	2	1	3	No	No need to restrict ODD – use permutation 6 instead
10	-	2	2	1	No	No failures, so use full ODD and behaviours (permutation 1)
11	-	2	2	2	No	No need to restrict behaviour – use permutation 8 instead
12		2	2	3	No	No need to restrict ODD – use permutation 6 instead

Figure 1: ODDs, Behavioural Competency Definitions and MELs that are available for a fictitious system, and a table defining which combinations are permitted (note that this is for the system design rather than the deployment, and hence uses ODD rather than TOD)

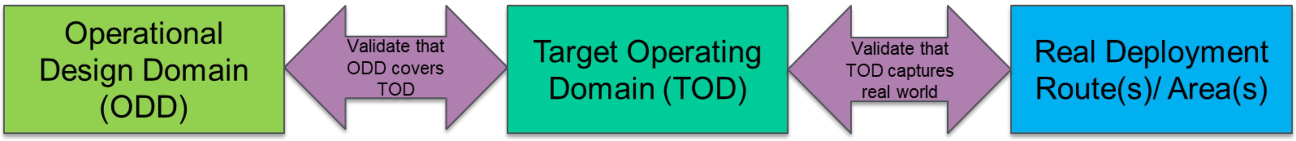

Operational Design Domain (ODD)
Validate that ODD covers TOD
Target Operating Domain (TOD)
Validate that TOD captures real world
Real Deployment Route(s)/ Area(s)

Figure 2: The ODD must, at a minimum, cover the scope of the TOD (i.e. the LSAV must be designed to cope with all the demands of the deployment). In turn, it must be confirmed that the TOD definition is an accurate reflection of the real world

Risk Framework and Safety Goals

A high-level safety analysis of road transport in general was undertaken to investigate hazards with the potential to harm road users, from which safety goals aimed at mitigating each hazard were formulated. A summary of the hazards considered is shown below:

- Collision between the automated vehicle and another object (moving or stationary);
- Direct harm to passengers from any of
    - motion of the automated vehicle (e.g., hard braking manoeuvre);
    - a moving mechanism on the automated vehicle (e.g., door mechanism);
    - technological hazards (e.g., electric shock, fire);
    - personal safety concern (e.g., medical emergency, assault on-board).

The aim of the risk framework is to require a level of functionality and performance that can be argued to provide an acceptable level of safety in the context of use. The overarching principles which underpin this risk framework and its objective acceptance criteria are that new automated driving technologies should:

- Not expose road users to unreasonable risk, and;
- Support the societal goal to make road transport safe for all road users.

The safety of the LSAV should be argued and demonstrated on a case-by-case basis for each system and deployment, rather than by setting a universal quantitative threshold for acceptable risk; the latter would be impractical given the diverse use cases, rapidly-evolving state of the art and shortage of existing empirical data for LSAVs. Manufacturers may demonstrate achievement of an appropriate level of safety by using risk acceptability principles such as ALARP (as low as reasonably practicable), GAMAB (*globalement au moins aussi bon, or 'overall, at least as good'*), and PRB (positive risk balance), and/or by using comparators such as an 'average' driver (as determined, for example, by road traffic accident statistics) or a 'competent and careful' driver.

However, it is recognised that is will not be practicable to generate statistically significant data to demonstrate quantitative equivalence to existing road traffic statistics until after the LSAV service has commenced. Therefore, the pre-deployment approval should apply reasonable due diligence to assure that appropriately robust engineering practice has been applied and that the resulting LSAV behaviour does not present unreasonable risk, whereas in-service monitoring should be used to collate statistics for the overall safety, thereby facilitating validation of the models and assumptions used within the pre-deployment approval.

The following top-level safety goals were formulated, which in turn should form the basis for technical performance requirements.

Safety Goal (1)	Do not cause collisions
Safety Goal (2)	Avoid foreseeable collisions
Safety Goal (3)	Protect all persons within and in the vicinity of the vehicle from harm

Assurance of the ADS (Automated Driving System)

Functional safety, safety of the intended functionality (SOTIF) and cybersecurity are essential elements in assuring the safety of LSAVs. For each of these disciplines, regulations and/or standards exist that capture industry best practice. This report therefore focuses upon what is required to augment existing regulations and standards such that any gaps pertinent to LSAVs are addressed, rather than attempting to duplicate them.

External inputs which support or influence the dynamic driving task, such as GNSS (global navigation satellite system) data or communications from an operations centre, pose a particular threat due to the potential for loss or corruption, either accidentally or via malicious actions, in turn yielding hazardous behaviours. Consequently, it is recommended that if the system makes use of wireless communications data, it must be able to maintain safe operation even in the event that these signals are lost or corrupted. Whilst these signals may act as inputs to support the ADS in its decision making, a cautious approach should be adopted with regards to allowing remote inputs to directly control the vehicle's motion.

Where machine learning (ML) is utilised within the ADS, a robust engineering methodology will be needed to provide assurance of this critical aspect of ADS design. Requirements for the subsystem performance should be specified, and the training data used within the development should be audited to confirm they are appropriate to meet these requirements (including consideration of the relevance, completeness, accuracy and balance of the data). Furthermore, the test data generated to verify the requirements should be similarly audited. Where redundancy is provided by other, non-ML components, this may reduce the assurance burden on ML components, which should be reflected in the allocated safety requirements.

Development of an appropriate architecture and selection of appropriate hyperparameters to control the learning process (such as learning rate and batch size) should be evidenced; this is likely to be an iterative process whereby multiple approaches are assessed and optimised via testing.

The requirements for ML-based subsystems should include quantitative metrics for performance, such as required levels of sensitivity and specificity of object classification, the accuracy of estimates for an object's position and speed, and robustness against variations within the operating environment. The testing should evidence that such performance metrics are satisfied. ML-based subsystems should also be tested once integrated into the full vehicle, to ensure that no undesired emergent behaviours become apparent. In-service monitoring should be used to validate assumptions about the system and its operating environment. A summary of the ML assurance activities are shown in Figure 3.

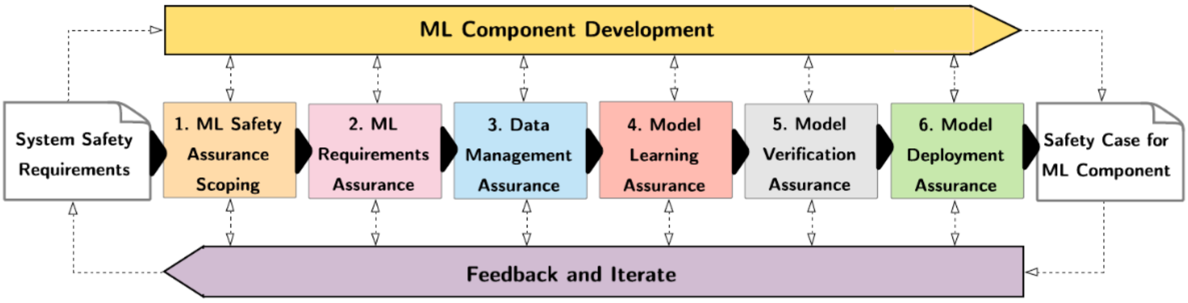

ML Component Development
System Safety Requirements
1. ML Safety Assurance Scoping
2. ML Requirements Assurance
3. Data Management Assurance
4. Model Learning Assurance
5. Model Verification Assurance
6. Model Deployment Assurance
Safety Case for ML Component
Feedback and Iterate

Figure 3: Machine learning safety assurance activities

Whilst the above ML recommendations are applicable to supervised machine learning trained via datasets, an alternative approach is reinforcement learning (RL), where the artificial neural network optimises its functionality in use according to a reward function that determines the quality of the outputs. Where RL is used, requirements should be specified and verified as above, but rather than auditing training data, it is instead necessary to validate the appropriateness of the reward function under all conditions; a naïve reward function can result in 'reward hacking' where the system optimises itself in an undesired manner.

'Offline RL', where the learning is completed prior to approval and the system then remains 'frozen' in deployment, is permissible. However, 'Online RL', where the system continues to learn in service such that the behaviour for each ADS continually changes, should be prohibited unless it can be clearly demonstrated that a failure of the function does not have an impact on safety or that changes to the previously assured function during operation can be restricted to safe ranges that have been pre-determined and validated as part of system development activities.

Test evidence derived from multiple test modalities will underpin the safety argument made in the safety case report. Further to traditional requirement verification for the components and (integrated) subsystems, the evidence must also include whole-vehicle test outcomes within realistic and TOD-representative scenarios that the deployed LSAV may be expected to encounter ('scenario-based testing'). Evidence must be provided to support the assertion that a sufficiently large and well-distributed sample has been taken across the space of reasonably foreseeable events and their combinations ('sample size' and 'sample coverage'). This must include justification of a validated test selection and generation methodology, and must sample from the range of behavioural competencies the vehicle is required to perform and the range of environments it must do so within (Figure 4).

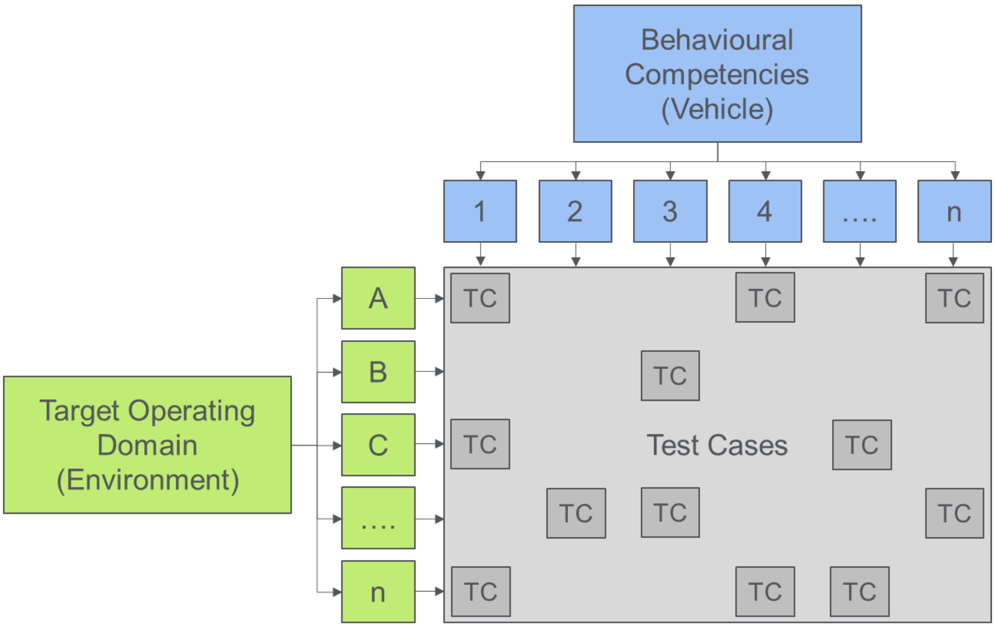

Behavioural Competencies (Vehicle)
1
2
3
4
....
n
Target Operating Domain (Environment)
A
B
C
....
n
Test Cases
TC

Figure 4: Schematic illustration of how the test cases selected for the system need to sample not just the range of Target Operating Domain (TOD) permutations that the ASDE could encounter, but also the range of behaviours that the system will be required to perform within the TOD

It is expected that the test programme would include a combination of mileage accumulation in the real world (which will be effective at exposing the system to more commonplace scenario permutations with perfect realism) and designed test cases in simulation or in a controlled physical environment such as a proving ground (which will allow coverage to be built up across the wider 'scenario space', including edge cases and hazardous events, but may introduce inaccuracies due to the difficulty of modelling the complexity of the real world). However, it is not proposed that regulatory requirements should be set for the proportion of each type of testing, or for a minimum mileage or duration requirement to be imposed; instead, flexibility should be allowed for the manufacturer to identify a suitable approach and argue its sufficiency for the particular application.

If it is to be used as evidence to support the safety case, any test data not collected in the real world, such as simulation (e.g. Software-in-the-Loop, Hardware-in-the-Loop) or mock-ups of a real scene on a proving ground, must be validated via a comparison of the correlation against real-world test results for the same scenario. The argument for acceptable correlation presented within the safety case should consider the quality of the models used and the similarity of the overall vehicle-level data. This could, for example, use metrics such as position and speed of the LSAV, attributes of any objects detected and classified by the system, radar cross sections observed by the sensors etc.

Assessment of the results of any test should extend beyond crude consideration of whether a collision occurred, to instead include:

- Driving context; for example, there will be some scenarios where a collision is unavoidable due to the actions of other road users, but the system should be judged to have performed well if the level of mitigation (e.g. by emergency braking) compares favourably with what a human driver would likely have achieved. Depending upon the overall risk framework approach selected, this comparison could be based on characteristic benchmarks such as a 'competent and careful' driver, or an 'average' driver, derived from existing road traffic datasets.
- Potential false-negative leading indicators; test programmes could reasonably be made more efficient at identifying system flaws, compared against cruder methodologies that look just at the 'global' outcomes, by flagging and reporting any *circumstantially inconsequential* failures. For

example, the failure by a vehicle to detect a pedestrian, who just so happens not to enter directly into the path of said vehicle, should be recorded and noted as a 'near-miss' style failure which had the potential to yield harm, but transpired on that particular occasion not to do so. This implicitly reduces the test burden since 'fewer stars need to align' for a system flaw to be uncovered.

Reaching a decision upon the overall acceptability of the LSAV requires aggregation of all the evidence generated from individual test cases. This should be a function of: the quality of coverage; the proportion of scenarios in which the performance was deemed to be acceptable; the fidelity of each test modality; and the inherent, residual risk presented by any 'failed' scenarios, which itself derives from both the severity of the consequences and the level of exposure to the identified triggering conditions. This evidence aggregation will ultimately lead to an assertion that the vehicle is shown to present reasonable or unreasonable risk during operation *with a certain level of statistical confidence*; this must then be judged by the appropriate stakeholders (e.g. regulators, insurers) as to whether it is *certain enough*.

Human Factors

Human factors should be given consideration within both safety case reports, including assurance that all persons who interact with the vehicle have an appropriately clear interface with which to interact safely and confidently. This includes employees such as customer assistants or maintenance staff, as well as passengers and other road users. The safety case reports should include consideration of normal operating conditions but also of emergency situations; in the latter, people may be in a state of distress and unable to think clearly, and interactions with emergency service or vehicle recovery personnel will be vital. Inclusive design principles, to ensure people with a range of additional needs are able to interact with the vehicle successfully, should be applied to help ensure a positive and safe user experience for all, and to help maximise the societal mobility benefits.

Particular consideration should be given where remote assistants play a safety-critical role in providing help to the vehicle when it needs guidance on the next appropriate move; such assistants must be provided with the necessary interface to allow adequate situational awareness and control. Any assumptions or estimates relating to the ability of humans to make timely and appropriate inputs to the system must be subjected to a thorough analysis of potential failure modes and must be validated through testing.

Operational Safety of the Deployment

A safety case report must be submitted by the Operator to demonstrate that the deployment will be acceptably safe. This must include consideration of any hazards that are specific to the route(s) or geofenced area(s) of the deployment, as well as how workshop procedures such as repairs and routine maintenance will be managed so that the vehicle continues to meet the manufacturer's specification.

A risk assessment should be included to prioritise the operational hazards for mitigation, and all resulting mitigations should be recorded. Consideration of relative risk in comparison to a baseline such as existing road traffic collisions may be informative in identifying hazards and estimating the scale risk they present (in line with the GAMAB or PRB paradigms referred to previously). However, it should be borne in mind that ALARP (as low as reasonably practicable) and the similar SFAIRP (so far as is reasonably practicable) are the test for the acceptability of risk management used by the UK's Health and Safety Executive (HSE) and the UK courts. Therefore, regardless of other methods used to support the operational safety risk assessment, it should be ensured that the deployment is in accordance with the principle of ALARP and/ or SFAIRP.

It is unlikely to be feasible for every member of staff with a safety-critical role to read, understand and recall the entire contents of the deployment safety case. Therefore, it must be ensured that accessible documents are in place such that each member of staff has access to a 'single source of the truth' that provides a clear understanding of the safety procedures and responsibilities required as part of their role.

In the absence of full commercial, large-scale deployments of LSAVs at the time of writing, existing standards and guidance relating to automated vehicle trials (such as the Safety Case Framework published by Zenzic, or BSI PAS 1881 and 1884), and documents relating to operational safety in other industries such as rail, should be used as the best available benchmark for good practice. However, once commercial deployments commence, and particularly once they begin to scale up, it is expected that a more mature state of the art will emerge. Regulators should therefore have a process to collect data on operational safety, and to use this to progressively develop LSAV operational safety regulations.

Safety Management Systems

Both the Manufacturer and the Operator should provide evidence within their respective safety case reports that they have an effective safety management system (SMS) in place during the development and operation of the LSAV. The SMS should be bespoke to the system and deployment, and should follow the 'plan, do, check, act' process during its creation and improvement. A safety policy must be established to capture the organisation's commitment to safety and to the SMS implementation, including the responsibility for senior staff to help foster a strong 'safety culture'.

The SMS should define processes through which in-service data will be collected and used to trigger improvements, and set out how staff will be selected, trained, assessed and managed. For any safety-critical roles, this should include consideration of what expectations can reasonably be placed upon humans in the role and how to mitigate human error, such as the implementation of a fatigue risk management system.

Safety objectives and safety performance indicators should be defined to provide benchmarks against which in-service safety data can be compared. Safety risk assessment, safety reporting and employee consultation should also form a key part of a robust SMS and strong organisational safety culture.