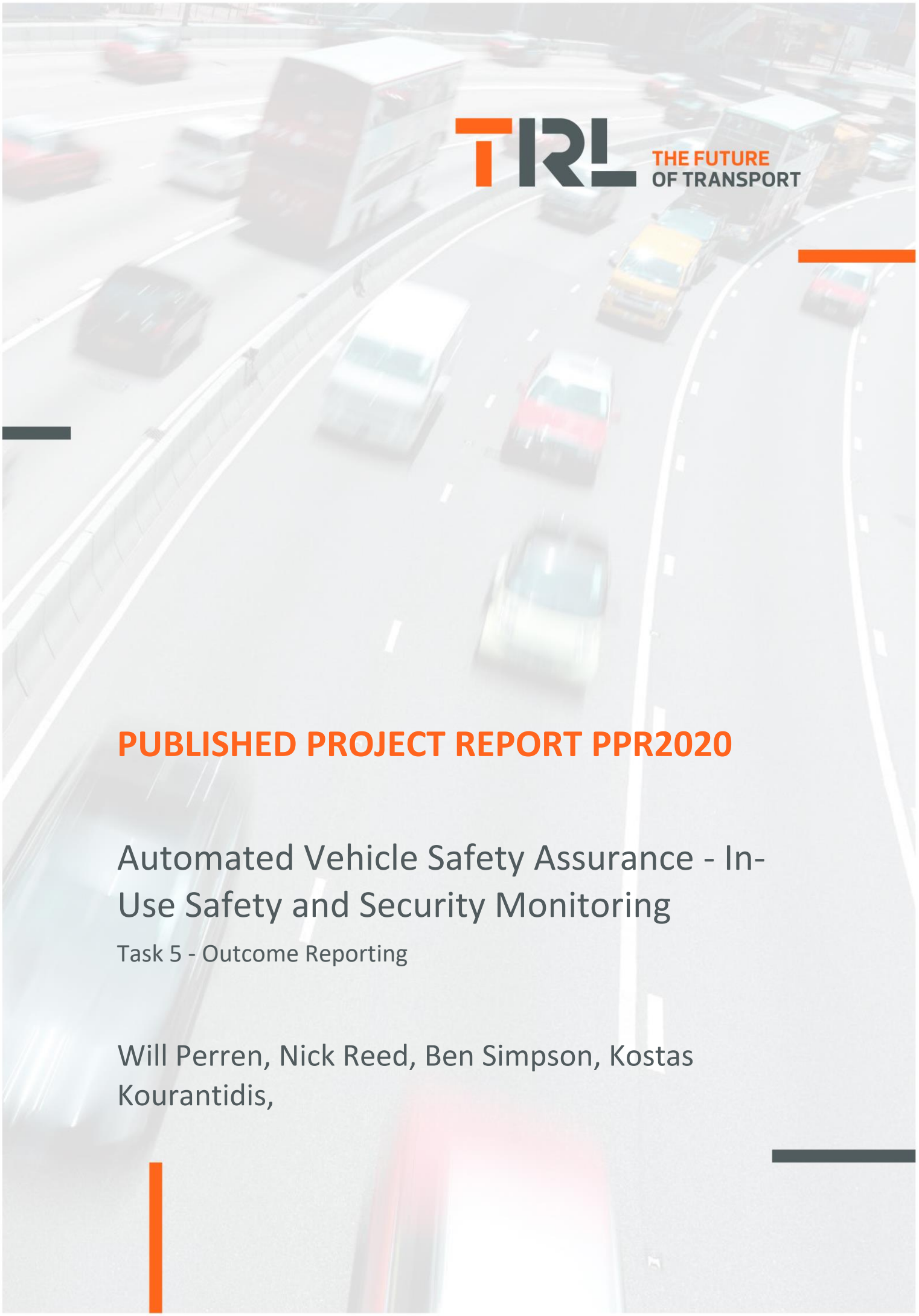

THE FUTURE OF TRANSPORT

PUBLISHED PROJECT REPORT PPR2020

Automated Vehicle Safety Assurance - In-Use Safety and Security Monitoring

Task 5 - Outcome Reporting

Will Perren, Nick Reed, Ben Simpson, Kostas Kourantidis,

Report details

Report prepared for:	Department for Transport		
Project/customer reference:	TETI0042		
Copyright:	© TRL Limited		
Report date:	30/06/2022		
Report status/version:	1.0		
Quality approval:			
Gareth Slocombe (Project Manager)	G. Slocombe	David Hynd (Technical Reviewer)	D. Hynd

Disclaimer

Executive Summary

The joint Law Commissions' report on automated vehicles outlined numerous methods for regulators to assess and report safety performance in line with a defined safety standard. Their key recommendation in this area was that Ministers should set an appropriate safety standard for Automated Vehicles (AVs) with support from experts (Law Commission and Scottish Law Commission, 2022). No matter how the safety standard is developed, the proposed In-Use Regulator will be expected to develop practical ways of measuring and reporting on current safety performance against the standard.

In this report, we suggest that a single measure of performance based on traffic safety statistics is not sufficient. Instead, a combination of both leading and lagging indicators of safety performance are required that monitor different elements of safety. This report evaluates a series of both leading and lagging Safety Performance Indicators (SPIs) that may be used to evaluate safety performance.

Collection of leading SPIs allows for much larger datasets to be generated compared to the data available from solely measuring lagging SPIs such as collision rate. However, while leading SPIs their correlation to an increased exposure to risk is yet to be established. This cannot be known prior to deployment. As such, it is necessary for some data to be collected as soon as possible so that the process for outcome reporting can be refined over time.

For the the proposed SPIs to be meaningful in any way, they must account for the variables that affect risk exposure and it is likely that different operational design domains (ODDs), deployment contexts and use cases may require different safety performance targets to assess the SPIs against. As such, methods are proposed to segment data by key variables known (to the best of our current knowledge of AVs) to have an impact on risk exposure. For each of these variables, the value of capturing them is assessed alongside the availability of the data required to capture them effectively.

Early AV deployments will accumulate far fewer miles than conventional vehicles in the same time period. As such, the rate of occurrence of risk events cannot be directly compared between automated driving systems and conventional driving. Methods for normalising data sets are required that takes into account variables that affect an AV operation's exposure to risk. Guidance is required to provide manufacturers with a consistent method of how to present safety performance data which should align where possible with international approaches to establish interoperability and access to much larger datasets in time.

The proposed in-use monitoring scheme primarily relies on the capture of in-vehicle data to monitor AV safety, which is a key input of aggregated data analysis for monitoring safety performance. However, the use of other data sources, such as incident investigation data, public incident reports and police reports cannot be overlooked as these provide coverage of some safety relevant events that the vehicle may not or can not detect itself.

There is a need to assess safety performance against a defined standard, with conventional human driving as the preferred reference by the Law Commission of England and Wales and the Scottish Law Commission (and those consulted by them). To compare with the wealth of data collected by an AV, equivalent datasets for leading SPIs and risk variables need to be collected. In practice, this means human driving performance will need to baselined in

`

comparable scenarios, use cases and deployment contexts. A naturalistic driving study with a methodology consistent with the data required in the in-use monitoring scheme and/or data from real fleets with cameras and advanced telematics would likely be the best way to collect this data.

The findings of this work have been summarised into a high-level process for aggregated data analysis which is shown in Figure 1. The associated process steps have then been assigned to different stakeholders through the use of a RACI (Responsibility, Accountability, Consulted, Informed) matrix to highlight how each stakeholder is involved in the process. This is shown in Table 1, accompanying the process flow.

`

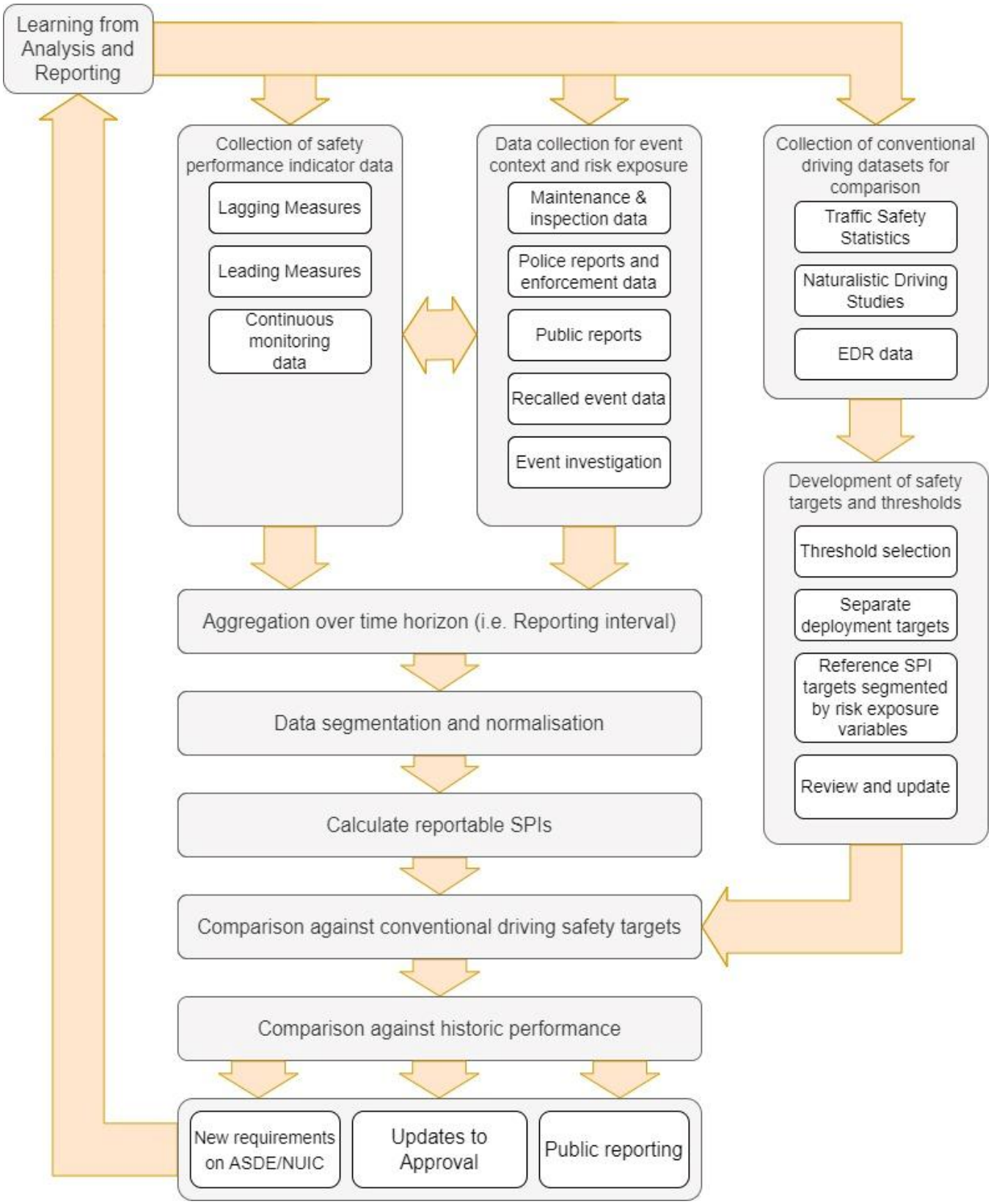

Figure 1: Proposed safety performance reporting process

`

Table 1: RACI Matrix for the proposed safety performance reporting process

	In-Use Regulator[1]	Approval Authority[2]	Manufacturer	Operator
Collection of safety performance indicator data	I		A	R
Data collection for event context and risk exposure	I		A	R
Aggregation over time horizon (i.e., Reporting interval)	I		A	
Data segmentation and normalisation	I		A/R	C
Calculate reportable SPIs	I		A/R	C
Collection of conventional driving datasets for comparison	A/R	C	C	I
Development of safety targets and thresholds	R	A	C	I
Comparison against conventional driving safety targets[3]	A/R	I	I	I
Comparison against historic performance[4]	A/R	C	I	I
Placing new requirements on manufacturer/operator, Updates to type approval, public reporting	R	A	C	C

The purpose of this reporting process is threefold.

- To feedback on safety performance to the manufacturer, in order to assess compliance with the safety case submitted at approval (and other scheme requirements) and set out any necessary remedial action
- To generate learnings around limitations of an approval scheme, that need to be addressed; and
- To report on the safety performance of automated vehicles in GB for the purposes of generating industry wide knowledge and assure the public that there is sufficient and robust oversight of these vehicles.

[1] In-use regulator also includes any investigating bodies.

[2] And Authorisation Agency, as appropriate

[3] This comparison would be provided only to the manufacturer it relates to. Scheme-wide statistics would be produced and which are shareable with public and wider industry

[4] This comparison would be provided only to the manufacturer it relates to. Scheme-wide statistics would be produced and which are shareable with public and wider industry

`

Table of Contents

List of tables and figures

Tables

Figures

`

List of abbreviations

ADS:	Automated Driving System
AV:	Automated Vehicle
CAV:	Connected and Automated Vehicle
DQF:	Data Quality Framework
DSSAD:	Data Storage Systems for Automated Driving
EDR:	Event Data Recorder
GB:	Great Britain
GIDAS:	German In-Depth Accident Study
LSAV:	Low Speed Automated Vehicle
MRC:	Minimum Risk Condition
MRM:	Minimum Risk Manoeuvre
NDS:	Naturalistic Driving Study
NHTSA:	National Highway Traffic Safety Administration
ODD:	Operation Design Domain
OEDR:	Object and Event Detection and Response
ORR:	Office of Road and Rail
PET:	Post Encroachment Time
RACI:	Responsibility, Accountability, Consulted, Informed
RAIDS:	Road Accident In-Depth Studies
RSS:	Responsibility Sensitive Safety
SPI:	Safety Performance Indicator
TTC:	Time To Collision
WP:	Work Package

`

1 Introduction

One of the most common and well-known road safety statistics is that human error is a factor in almost all road traffic collisions. In Great Britain in 2020, driver/rider error contributed to 65% of all collisions, impairment, and distraction (including drugs and alcohol) contributed to 17% and behaviour and inexperience contributed to 25%[5]. These statistics are often used as justification by industry and government to push for the development and introduction of Automated Vehicles (AVs), because AVs in theory shall never be able to make these mistakes.

In this context, it is often claimed that AVs will be safer than human drivers because they have the potential to eliminate collisions where human driver error is a factor, thus reducing the impact on public health and the associated social and economic costs. However, alongside these potential benefits, the introduction of AVs may also bring about potential risks. Factors such as inclement weather, complex driving tasks and unforeseen situations are known to be difficult for AVs and poor AV performance may increase the safety risk. New technologies also bring the possibility of new types of collision risk factors. For AVs, this could include programming errors, sensor faults or unforeseen behaviours by the AVs.

Public confidence and, ultimately, acceptance of AV technology will depend on whether AVs are truly safer than conventional driving (Kyriakidis *et al.*, 2015). While it has been shown that public confidence is increased by exposure to positive experiences with the technology (Penmetsa *et al.*, 2019), negative experiences - such as a collision involving an AV - are more likely to come under greater public and media scrutiny than a collision involving human drivers only. In these circumstances the claims made around the safety benefits of AVs are likely to come into question. As such, the safety benefits of AVs are a key open question. So far, there is insufficient evidence to demonstrate with statistical significance the safety benefits of AVs compared to conventional driving; validation of this claim will be vital to acceptance and uptake of the technology by consumer and society at large.

It has been reported that traditional validation of safety performance prior to deployment (i.e., accumulation of fault-free miles driven in the real world above an acceptable threshold) is largely unfeasible for AVs (Kalra and Paddock, 2016). As such, robust claims about safety performance compared to human driving can only be made following deployment. The introduction of AVs in the UK will provide the opportunity to start to collect AV safety performance data and measure their actual safety benefit.

[5] Includes only accidents where a police officer attended the scene and in which a contributory factor was reported. https://www.gov.uk/government/statistical-data-sets/reported-road-accidents-vehicles-and-casualties-tables-for-great-britain

`

2 Scope and Purpose

This report explores how in-use monitoring data collected to validate AVs against approval requirements during operation can also serve as a basis for the statistical evaluation of wider AV safety performance compared to human driving. It is expected that an acceptable target for AV performance will need to be set in relation to current human driving performance. As recommended by the Law Commission (Law Commission and Scottish Law Commission, 2022), setting a safety standard in line with public expectation is thought to be a political decision. As such, this report does not seek to develop a safety target or standard. Instead, the report explores how an in-use monitoring framework developed for the GB AV approval scheme can incorporate a mechanism for collecting data and assessing AV performance under the scheme to support measurement against the safety standard when developed.

This report discusses method to evaluate safety performance during operation. In previous work under Work Package 5 (WP5), data and measures have been defined for the purposes of identifying events that potentially have safety relevance and data to provide context and understanding of the event. For this, measures have been defined split into leading and lagging measures.

Lagging measures specifically target data capture from extreme trigger events, which highly correlate to adverse risk outcomes (e.g. typical severe collision scenarios). They give insight on events that already occurred which are typically small in number and, as such, without a clear ability to estimate future likelihoods given low occurrence. Leading measures target data capture of vehicle operations that have the potential to become realised risk events. They are a proxy for actual risk occurrence. They give insight into potential risk and are typically much larger in number than lagging measures.

These measures focus on identifying and qualifying single events that may indicate potential non-compliance with approval requirements, safety case arguments or traffic rules. In doing so, a recommended data set has been defined to support the identification and subsequent analysis of unsafe events.

The focus of this report is to define measures that enable the statistical evaluation of safety performance. For clarity these have been termed Safety Performance Indicators (SPIs) in this report. SPIs are not the same as leading and lagging measures; they are derived from them. As such SPIs can also be leading or lagging. Leading and lagging measures are individual-level data, meaning they represent a single observation (i.e. data point). To enable statistical evaluation of safety performance, leading and lagging measure data must be collected over time to produce aggregate level data. The data then has to be processed in order to remove false positives, separate by variables that affect safety, and normalised by exposure for a statistical comparison to be made. As such, SPIs can be seen aggregated and processed leading and lagging measures. For example, an example of a lagging measure is the existence of a collision between the AV and another vehicle. The corresponding lagging SPI would be the frequency of that type of collision per unit distance (e.g. in the last million miles miles). Recommendations are made in this report in how this process could work for the in-use monitoring scheme.

The SPIs have been developed on the basis of the data expected to be available in the scheme (i.e. the already defined minimum dataset specification (Chapman and Perren, 2021)).

`

However, every attempt is made to conceive of additional data necessary to allow or enhance the analysis. The value of this additional data for the analysis, as well as the practical issues with collecting it are discussed, to provide a balanced recommendation.

`

3 Requirements for a Safety Standard

The Law Commission of England and Wales and the Scottish Law Commission, in their joint report and their consultation papers before it, discuss the subject of automated vehicle (AV) safety and what it means to be safe enough to be permitted on roads or other public places. (Law Commission and Scottish Law Commission, 2022)

In its consultation, the Law Commission asked consultees the question of "how safe is safe enough" for AVs, offering three possible standards:

(a) As safe as a competent and careful human driver (option A).

(b) As safe as a human driver who does not cause a fault incident (option B).

(c) Overall, safer than the average human driver (option C).

The key recommendations from the Law Commission's report regarding the safety standard for AVs can be summarised below:

- The safety standard must be measurable
- The decision over how safe an AV should be depends on whether the remaining risk is acceptable to the public. This is essentially a political and not a technical decision, which means the safety benefits must be demonstrable
- There must be equity in the distribution of remaining risk; AVs, while safer overall, should not be more dangerous to one road user group vs another, e.g. safer for motorised users but more dangerous for pedestrians. This aligns with recommendations to the European Commission on Connected and Automated Vehicle (CAV) ethics (European Commission, 2020)

The recommendations of the Law Commissions' report have a clear impact on how the safety standard will be set, as comparisons between the performance of Automated Driving Systems (ADS) and "conventional" (manual) human driving are key. This in turn, will have an impact on the data types used to determine the safety standard (data requirements) as well as the mechanism through which those data will be collected and stored while the vehicle is in-use.

A major limitation for comparing AVs and human drivers and setting a safety standard based on the comparison is the lack of comparable data for existing conventional vehicles. While there is relevant research around human driving behaviour and collisions or near miss events (for example, the NHTSA 100 Car Naturalistic Driving Study (National Highway Traffic Safety Administration, 2006)), there is no agreed systematic approach to collecting and processing this data at scale. Defining the measures, dataset, and framework for collecting data is required; the collision and near miss reporting requirements already defined provides a structure for this for automated vehicles during operation (Balcombe and Perren, 2022). For a comparison with conventional vehicles, equivalent datasets are required (see Section 5.5).

As such, no matter how the safety standard is set, the data needed to measure it in-use will need to be:

- Comparable to conventional driving datasets, either through existing data or through further research.

`

- Able to measure risk exposure to different road user groups.
- Not solely based on traffic collision statistics
- Intuitive and comprehendible by the public, policymakers, and regulators.

These requirements form the basis of the discussion on monitoring safety performance in this report.

`

4 Measuring Against the Safety Standard

4.1 Safety Performance Indicators

Measuring the safety performance of AVs and comparison with data relating to human drivers is essential in setting the safety standard. The in-use regulator also requires the collection of such data from the vehicle to continuously monitor AV performance and reassure the public about the safety of AV technology.

A comprehensive list of safety relevant events of interest can be found in the WP5 Task 1 report – In-Use Monitoring Taxonomy (Reed, 2022). The in-use monitoring scheme should broadly be able to collect information on:

- **Collision events** (lagging measures)
- **Near miss events** (leading measures)
- **Safety-related violations,** such as exceeding the speed limit; running red lights; or careless or dangerous driving (also leading measures)
- Other **road rule violations**, not directly related to safety but that negatively impact the flow of traffic or safe movement of other road users.

Data collected surrounding these events provide the opportunity to identify the Safety Performance Indicators (SPIs) that are most useful in measuring the safety performance of AVs.

4.1.1 Lagging SPIs

Lagging SPIs are the indicators that measure actual harm (crashes) and their outcomes (RAND, 2018). They give insight on events that are typically low in number (as the number of collisions is generally low compared to the total miles travelled by a vehicle fleet) and therefore without clear ability to estimate the future likelihood of an event. These data measures are used to flag when a collision or actualised risk outcome has happened and help investigate individual events.

Traffic collision research investigates collisions and gathers data that can help understand collisions, how and why they happened. The Road Accident In-Depth Studies (RAIDS) project in the UK (Cuerden and McCarthy, 2016), managed by TRL on behalf of the Department of Transport; the Reported road casualties Great Britain, which uses the STATS19 reporting system, (HM Government, 2021); and the GIDAS (German In-Depth Accident Study) (GIDAS, 2019) project are examples of such databases that contain collision data. In the UK, RAIDS and STATS19 hold information about collisions such as time and location, road conditions, vehicle details and manoeuvres, casualty/injury details and contributory factors to the collision. Generally, the STATS19 database contains high-level information on more collisions while RAIDS contains a smaller number of collisions but in much greater detail. This data can then be used to evaluate road safety by calculating different traffic collision statistics.

Previous work under WP5 recommended series of data sets which are required to identify possible collision and injury events for low-speed automated vehicles (LSAVs) (Chapman and Perren, 2021). These lagging measures are used to trigger data capture surrounding an event

that may potentially be of interest. While the proposed data sets of demonstrate a good ability to detect events, they do not directly allow measurement of collision risk over time in this current state. A degree of processing is required to convert them into comparable SPIs relating to collision risk. Firstly, the presence of an actual collision following any of these triggers needs to be identified. For example, a safety envelope close proximity trigger does not distinguish as to whether a pedestrian was struck by the vehicle or simply brushed passed it. As such a degree of investigation is required to verify the event and understand severity and context (see Section 5.3). Following this investigation, the relevant event information can be coded and filed for aggregated analysis.

It is also evident that there is not complete compatibility between the recommended minimum data sets and data stored in collision databases such as STATS19 or RAIDS. The reason for this is twofold. Firstly, not all of the data previously proposed in the WP are recorded by these databases. Secondly, as these databases only record collisions (not other risk events associated with AV operation that could lead to casualty or injury). The proposed data sets for AV performance are expected to be available as defined in the minimum data set specification for this project (Chapman and Perren, 2021)

For measuring human performance of the same SPIs for comparison (see Section 5.5) will require further data collection. Useful data are usually part of research projects, such as the NHTSA 100 NDS, but are limited in number, scope, and relevance. The relative scarcity of such data is an issue for comparability between conventional and AV driving; one that may partially be addressed if more EDR-equipped vehicles are on the road. For example, EU requires new vehicle types of the M_1 and N_1 categories sold in the bloc after July 2022 to come equipped with EDR devices. (European Commission, 2022). This will not enable capture of like-for like lagging measures but would allow identification of comparable collisions and injury events used to aggregate Lagging SPIs for human driving. It could also partially generate similar data to categorise events to allow comparison of Lagging SPIs between AVs and conventional driving.

Lagging SPIs are characterised by the following:

- High correlation to risk incidents that have actually happened and can be measured (precision).
- Reasonable coverage of a risk incident (recall) and its severity. This includes lower severity incidents; however, may result in a drop in recall as some events may not be recorded by SPIs.
- Data can be lost in rare instances where the data recording mechanisms or triggering sensors are damaged (severe collisions) which may skew the findings towards lower severity events.
- The low occurrence of collisions with traffic fatalities and injuries (compared to vehicle miles travelled) indicates that it would take a significant amount of time to build a proper sample of AV collision data and compare automated and conventional driving in a statistically meaningful way (Kalra and Paddock, 2016). Using lagging SPIs and relevant data without building confidence first may lead to loss of trust by the public which would be harmful to the AV industry.

`

- Given the low occurrence of lagging SPIs, the perceived frequency in AV system updates (which may drastically alter) would likely not allow for data on enough collisions to be collected to measure safety performance between updates, though this may provide insights into how vehicle safety performance is improved over time.

Kalra and Paddock (2016) attempted to estimate how many miles AVs would need to drive (and how much time that would take) to demonstrate a certain level of safety. Using current US collision statistics for conventional vehicles (1.09 fatalities/100 million miles) and making assumptions regarding statistical confidence (95% confidence level) and AV fleet operation (100 vehicles driving 24/7 at an average speed of 25mph) they calculate that, to demonstrate that the same fatality rate is achieved as conventional driving, AVs would need to drive 275 million miles **fault-free**, which would take the fleet of 100 AVs over **twelve years** to achieve.

A similar simple calculation can be attempted for the AVs being deployed in the UK using UK collision data (DfT, 2020) and LSAV characteristics while also making some essential assumptions. The following calculation highlights the required miles driven to demonstrate that AVs are as safe, or safer, than conventionally driven vehicles. In a scenario whereby AVs were involved in substantially more collisions than conventionally driven vehicles, then a conclusion that AVs are less safe than human drivers could be reached in a shorter timespan.

According to DfT's 2019 annual report on road casualties (pre-covid data), the fatality rate per billion vehicle miles was 4.87 fatalities per billion miles. However, that figure includes all fatalities across the whole GB road network. The number of fatalities on urban roads (where early deployments of automated vehicles are foreseen) is 653, with traffic estimated at 135 billion miles for urban roads. The rate of fatalities for urban roads is therefore 4.84, only marginally lower than the national average. It should be noted that urban roads include both urban 'A' roads and minor roads, with varying speed limits often in excess of a Phase 1 vehicle max speed.

To demonstrate that LSAVs have a fatality rate of 4.84 per billion miles with a 95% confidence level, the vehicles would have to drive approximately **612 million failure-free miles**. With a fleet of 100 vehicles being tested 24h per day, 365 days a year at a speed of 20mph, this would take almost **35 years.**

Assuming an average operating speed of 15 mph, the time required to collect the sample would be about 46.5 years.

Estimating the fatality rate of the fleet within 20% of the assumed rate (and 95% confidence for the standard distribution), the required miles are almost 20 billion (approx. 1,132 years to collect at a maximum speed of 20mph)

Clearly, it is impossible to collect this amount of data. This highlights the need for simulation and scenario-based testing prior to deployment, but to continuously validate against this target in-use would be impossible without taking a different approach. Leading SPIs are needed to widen the sample size of the data available. Furthermore, new methods assessing risk exposure, outside of accumulating vehicle miles are required. However, there is still value in collecting lagging measures as a means of both monitoring the change in safety performance over time and for establishing the predictive value of leading SPIs (as discussed in Section 4.3). Indeed, just as higher frequency leading SPIs can provide insight into the

expected frequency of collisions, lower severity collisions can provide insight into the expected frequency of more severe collisions (including fatalities). As such, they do still have value for analysis

To summarise, safety performance for conventional driving is measured and reported using statistical data such as collision rate, collision severity, fatalities, and injuries etc. Lagging SPIs from in-vehicle data are valuable in credibly in identifying hazardous events (such as collisions or other events that have actually happened, but not sufficient to identify other safety events such as near misses or traffic infractions) and can also be used to infer a collision rate; these, along with other reporting sources such as police reports, on-site collision investigations and infrastructure data, can be used to determine the collision rate for AVs. However, the number of AVs currently on the road does not allow us to build a data sample sufficient for a meaningful (statistically significant) comparison with conventional vehicles. With the collision rate of AVs being a key SPI for the regulator to determine whether AVs are as safe (or safer, as the safety standard would require) as conventional vehicles, other methods of acquiring this information are necessary.

It is not feasible to measure AV safety performance in this way. For the purposes of regulatory oversight however, it is still recommended to monitor this as it is possible to draw statistically significant conclusions that an AV is less safe than this target quite quickly. Aside from this, Leading SPIs and identifying their correlation to safety outcomes, are the key to measuring AV performance. This is examined in the next section.

4.1.1.1 Recommendation*:*

Monitor a set of lagging SPIs for AVs through collecting in-vehicle data (lagging measures and telematics), police reports, insurance reports, collision investigation and infrastructure data. The following lagging SPIs are recommended:

- Collisions classified by severity
- Passenger injury; and
- Other realised hazards (e.g., fires, noxious gas releases).

Appropriate methods should be established for aggregating lagging SPIs (as discussed in Section 5.1) and segmenting data by risk exposure variables (Section 5.4)

4.1.2 Leading SPIs

Leading SPIs measure the prevalence of events that are precursors to realised outcomes. As no collision necessarily occurs, these events are not inherently unsafe. However, their presence indicates that the system may not be performing safely or as expected. As such, they can be used as proxies for collisions if a clear correlation can be established (RAND, 2018).

Leading SPIs have significant potential benefits. Firstly, they happen more frequently than their lagging counterparts. This enables faster learning and may allow for statistically significant results and useful SPIs to be obtained earlier in deployment (AVSC, 2021). Leading SPIs also enable learning without requiring dangerous events to come to fruition. This is a safety benefit, but is also beneficial because these events may be significantly detrimental to

the overall development of AVs due to their disproportional negative media attention and effects on public confidence in AVs.

However, there are two key challenges with using leading SPIs, based off of leading measures, for statistical evaluation of the safety of AVs. The first is that as yet, correlation to safety events has yet to be established and so their significance on safety cannot accurately be stated. The second is that they are difficult to define and are often highly contextual. This means that they are difficult to record in such a way that meaningfully differentiates them from safe driving behaviour.

4.1.2.1 SPIs considered

This section summarises a variety of proposed leading SPIs and assesses their usefulness in evaluating the safety performance of an AV. A review of literature was conducted and the following SPIs have been selected because it is believed that they are likely have a correlation to the safe functioning of AVs. Broadly, this belief is based on the existence of correlation to safety outcomes in human drivers (for example, acceleration profile) or based on the understanding of what is required for safe operation of a AV (for example, disengagements).

The leading SPIs can be broadly split into the following five categories:

- **Vehicle kinematics and driving style:** How the vehicle travels through its environment, such as acceleration profile.
- **Proximity measures:** How close the AV gets to other road users, and for what period of time. Close proximity gives smaller margins for error.
- **Internal vehicle health:** This includes SPIs such as correct object detection and how often the AV is no longer able to continue its journey in the event of a disengagement or ODD exit.
- **Appropriateness of behaviour:** This includes SPIs which involve an assessment of the vehicle's behaviour, such as whether the vehicle committed a traffic infraction or performed an appropriate action in a given situation.
- **User feedback:** How users, such as passengers, feel the AV has performed, and how safe they felt during its operation.

The leading SPIs considered during this work are presented below:

- **Acceleration of ego vehicle:** This would include longitudinal and lateral acceleration of the ego vehicle. Instantaneous high acceleration may indicate evasive action taken by the AV, while sustained periods of high acceleration may indicate improper driving style.
- **Jerk of ego vehicle:** Jerk is the rate of change of acceleration. High jerk may be indicative of constant corrective actions being taken, or poorly calibrated sensors.
- **Trigger of electronic stability control (ESC):** ESC systems use automatic braking of individual wheels to assist in maintaining control of the car in critical driving situations involving a loss of traction. This could be the result of a failure to assess road

conditions, leading to a loss of control and is indicative of driving beyond the vehicle's capabilities.[6]

- **Acceleration of surrounding vehicles:** unusual acceleration profiles of surrounding vehicles may indicate unsafe behaviour. In such an example, surrounding vehicles may be forced to take evasive action to avoid a collision with the AV.
- **Time-to-collision (TTC):** The time it will take for entities to collide, assuming their current velocities and direction are maintained. By definition, all collisions result in a TTC = 0. Periods of low TTC may indicate near misses while sustained low TTC may indicate unsafe driving behaviour. It should be noted that there are a number of mathematical variations of TTC. For example, modified time-to-collision (MTTC) assumes that the entities' current accelerations are also maintained. For the purposes of this assessment, these have all been considered under TTC. This is because their applications, and the expected requirements for implementing each, are very similar.
- **Safety envelope violations:** The safety envelope is the physical space around the ego vehicle, inside which a collision may not be avoidable if another entity is present. Of pertinence to measure is how often safety envelope is violated, by who, by how much, and how quickly restored. This is closely related to TTC (TTC is one measure of calculating the safety envelope) and may be indicative of tactical awareness and forward planning. Different algorithms have been developed for the assessment of a safety envelope in real-time have, such as Intel Mobileye's Responsibility Sensitive Safety (RSS) (Mobileye, 2020) and Nvidia's Safety Force Field (Nvidia, 2020).
- **Post-encroachment time (PET):** The time between one road user departing from a location of potential collision to the time another road user arrives in that same area. This is also referred to as following time or following distance. It is closely related to safety envelope and TTC.
- **Vehicle's correct detection of objects and instructions:** How often, and when, the AV correctly identifies and classifies hazards and other objects. This could also include the vehicle's ability to recognise and comply with visual instructions, whether they come from a person or from a sign. This is required for accurately perceiving the operating environment, which is key to safe operation of a vehicle. This may be based on adjustments to the classification of objects or other changes in perception; these could include discrepancies between the vehicle's predicted trajectory compared to its actual trajectory. It may also be based on operating a separate monitoring system (independent of the ADS). Both are considered as possibilities at this stage.
- **Disengagements:** The number of times the vehicle transitions from automated mode to manual mode during operation. This would include operator intervention. Disengagements during deployment are an indicator that something unexpected has happened and in which the vehicle may no longer be able to safely continue a journey.

[6] ESC, as well as other safety systems on the vehicle may or may not be present which means the collection of data relying on them may not be possible, unless such systems were mandated.

`

- **Ratio of disengagements to Minimum Risk Conditions (MRCs) achieved:** An example of an MRC is coming to a stop in a safe place on the side of the road. Failure to act appropriately in the event of a disengagement is likely to leave the vehicle in an unsafe position on the road.
- **ODD exit (or close to limit):** This would include surpassing the threshold of any element of the ODD as defined by PAS 1883 or similar standard. Exiting the ODD is not always due to a system fault, but may be indicative of poor planning from the vehicle, or an inability to accurately perceive the environment. Consistently operating near the limit of the vehicle's capability may cause failures more often.
- **Traffic rule violations:** For example, failure to stop before a line, crossing central markings, running a red light, or speeding. These violations often result in scenarios which are hazardous, irrespective of the entity controlling the vehicle. There are situations in which it is appropriate to commit minor violations, such as in order to provide room for an emergency vehicle. As such, AVs may occasionally intentionally break a rule. These instances may be differentiated from times when the vehicle unintentionally breaks a rule or commits and offence. **Proper response action:** Correct response taken within a specified threshold after an event occurs. This allows for more in-depth analysis of an event and may highlight instances of poor driving when not otherwise triggered by other measures.
- **Roadcraft:** Based on the 'Roadmanship' concept originally proposed by the RAND corporation (RAND, 2018)as a general measure of the vehicle's conduct. This may consist of a number of other indicators in this list (for example, acceleration, post-encroachment time, TTC). Overall, "good" road conduct is likely to result in fewer collisions and is a target for all road users.
- **Qualitative feedback regarding feeling of safety:** An overall measure of a other road users' feelings of safety while in the vehicle. Human beings tend to have good intuition for whether a situation was safe, and road users feeling unsafe is likely to be the result of poor driving.

A summary of the leading SPIs is shown in Table 2. The data used to aggregate these metrics are derived from the Leading and Lagging measures minimum data set specification (Chapman and Perren, 2021). Collection of this dataset would need to be confirmed at approval.

`

Table 2: Summary of leading SPIs

Category	Safety Performance Indicator
Vehicle kinematics and driving style	Acceleration of ego vehicle
	Jerk of ego vehicle
	Trigger of electronic stability control (ESC)
	Acceleration of surrounding vehicles
Proximity measures	Time to collision (TTC)
	Safety envelope violations
	Post-encroachment time (PET)
Measure of the vehicle's internal health	Vehicle's correct detection of objects and instructions.
	Disengagements
	Ratio of disengagements to MRCs achieved.
	ODD exit (or close to limit)
Appropriateness of behaviour	Traffic offences and Highway Code rule violations
	Proper response action
	Roadcraft
User feedback	Qualitative feedback regarding feeling of safety

4.1.2.2 Assessment of SPIs

In order to be able to make assessments of safety on the basis of leading SPIs, an in-use regulator will need access to information such as:

- The correlation of the SPI to lagging safety outcomes
- The feasibility of gathering data required to calculate the SPI
- How this compares to the benchmark used for determining acceptability, for example to the risk posed by human drivers

Given that leading SPIs can be present without a safety-related outcome, such as a collision, occurring, the fundamental criterion is whether they actually have a correlation to safety outcomes, and what the nature of this correlation is.

However, this will only be possible once AVs are in use and there is sufficient data available to draw such correlations. Until then, the way in which SPIs must currently be assessed is different. As such, other criteria which reflect these requirements have also been included: an overview of these and the justification for their selection is listed below.

- **Data availability:** For an SPI to be used, it must first be obtained. This criterion considers challenges relating to measuring and perceiving that an event has occurred. It also considers the reliability of sensors involved in obtaining the relevant data.
- **Expected correlation to safety-related outcomes:** Fundamentally, the question which must be answered for each criterion is whether it is indeed correlated to the lagging

safety outcomes which are to be avoided. While this is not currently attainable, it is possible to made educated estimations regarding a particular measure's correlation to safety-related outcomes.

- **Ease of defining a meaningful threshold which has correlation with safety outcomes:** It is very difficult to prescribe in a way which meaningfully differentiates between acceptable and unacceptable behaviour. Segregating data and defining thresholds relative to ODD, vehicle type, or use case, rather than across all use cases of automated vehicles may improve this. This is because a particular type of automated vehicle, such as an LSAV operating in a single, known environment is likely to have a much more tightly constrained window of expected performance, and will therefore be easier to establish if behaviour is outside of that window and therefore unexpected and/or unsafe. Methods of data segregation are explored further in Section 4.2.
- **Coverage of collision types and road users:** It is important to be able to evaluate the correlation between an SPI and all collision types (such as front collisions, side collisions) and for all road users, in order to be representative of all scenarios seen in the traffic ecosystem. This is particularly true where the AV is required to recognise some aspect of its own performance, for example, in disengagements and object and event detection and response.
- **Data points and sources required to calculate SPI:** With reference to the minimum dataset (MDS) specified in Task 2 (Chapman and Perren, 2021), additional vehicle data, and external data, this criterion highlights any gaps in what is required compared to what can be obtained.
- **Feasibility of gathering the required data points from the AV:** Related to the point above. For example, does it require supplementary information from connected infrastructure or other road users?
- **Ease of collecting comparable human data:** The Law Commissions' report (Law Commission and Scottish Law Commission, 2022) highlight the use of human performance as a benchmark for AV performance. brought findings of comparing against human drivers. This may prove challenging in some areas as not all SPIs, such as disengagements, have a human-comparable counterpart. For those that do, data is not recorded to the same degree as is specified for AVs. Human comparability is explored further in Section 5.5.
- **How intuitive it is, and can it be easily understood by the public:** Public acceptance is key to the adoption of AVs in the UK. In order to support this and educate the public on the benefits and risks of AVs, it is important that these SPIs used to determine that risk can be communicated to the public.

Each SPI was qualitatively assessed as "high", "medium", or "low" against each of the criteria summarised in Table 3.

Table 3 - Summary of assessment criteria for each SPI

Assessment criteria
Data availability
Correlation to safety outcomes
Ease of defining a meaningful threshold which has correlation with safety outcomes
Coverage of collision types and general operation
Data points and sources required to calculate SPI
Feasibility of gathering the required data points from the AV.
Ease of collecting comparable human data
How intuitive it is and can it be easily understood by the public

4.1.2.3 Recommendations

The following sets of recommendations outline the key criteria on which a decision should be based, and therefore the key SPIs which should be recorded. Two sets have been listed:

1. The ideal scenario, in which all the required information is readily available,
2. A scenario which can be actioned with currently available data and understanding of correlations to safety

`

This acknowledges the need to act now to address these questions around the safety of AVs, while providing a target for the industry to attain.

Recommendation set 1

In an ideal scenario, an assessment would be made on the basis of validated correlation to safety outcomes. This assessment would be made against a known benchmark, which is likely to include reference to human performance. This assessment would also be made on the basis of all collision types and for all road users. As such, the key criteria required to make a decision would be:

- Correlation to safety (validated)
- Comparability to human performance
- Coverage of event types

On the basis of this assessment criteria, it is estimated that the key SPIs which would inform such decisions would be:

- **Proximity data (TTC, safety envelope, PET):** These have a high correlation with safety in human drivers, and a high coverage of event types.
- **Vehicle's correct detection of objects and instructions:** Failures in human perception of a situation and other objects is highly correlated to collisions. However, a measure of this is difficult to develop as they represent unknown occurrences. This may however be used to provide context to other events (i.e., a factor in event causation – Section 5.3.6).
- **Proper response action:** By definition, if all road users took a proper response in each scenario, very few, if any, collisions would occur.
- **Roadcraft:** A combined measure would allow a singular measure to assess the quality of average driving behaviour by the HAV, which would ease interpretation by public and other stakeholders. However, this is dependent on all measures used in its calculation and their combination to be well understood.
- **Disengagements (including supporting information such as whether an MRC was achieved):** Disengagements are a useful indication that something unexpected has happened; analysis and post-processing of these events is likely to lead to significant learning.
- **Vehicle kinematic data (speed, acceleration, jerk):** These have a high correlation with safety in human drivers and are often indicative of near miss events such as evasive manoeuvres.
- **Traffic offences and Highway Code rule violations (or a subset thereof):** These have a high correlation with safety in human drivers. However, they are limited by the ability to detect a traffic infraction. It is suspected that some traffic infractions will not be detectable.

`

It should be noted that although validated correlation of safety would be leveraged in this ideal scenario, this recommendation list is based on estimated correlation with safety events. The method for establishing the predictive value of these measures through monitoring is discussed in Section 4.3.

Recommendation set 2 – IMMEDIATELY ACTIONABLE

The main challenges with currently assessing the safety of AVs are that correlations to safety outcomes have not yet been validated, and that it is not possible to accurately and consistently obtain the required SPIs. As such the key criteria for assessing which SPIs should be recorded in the first iteration of in-use monitoring are:

- Correlation to safety (expected)
- Data availability
- Data sources required, with respect to those specified in Task 2

On the basis of this assessment criteria, it is the key SPIs which may be recorded immediately and used as the basis for an initial evaluation of AV safety are:

- **Proximity data (TTC, safety envelope, PET):** These have a high correlation with safety in human drivers, and a high coverage of event types.
- **Disengagements (including supporting information such as whether an MRC was achieved):** Disengagements are a useful indication that something unexpected has happened; analysis and post-processing of these events is likely to lead to significant learning.
- **Vehicle kinematic data (speed, acceleration, jerk):** These have a high correlation with safety in human drivers and are often indicative of near miss events such as evasive manoeuvres.
- **Traffic offences and Highway Code rule violations (or a subset thereof):** These have a high correlation with safety in human drivers. However, they are limited by the ability to detect a traffic infraction. It is suspected that some traffic infractions will not be detectable.

4.2 Establishing Thresholds

For all measures, a threshold value must be defined such that any value that exceeds that threshold triggers recording of data required for aggregated analysis (as well as comprehensive data required for in-depth investigation). While the measure aligns to risk behaviours, ultimately the threshold is what delineates between what is considered acceptably safe and unacceptably safe performance.

Setting thresholds also play a crucial role in determining how well the events recorded by exceeding such thresholds correlate to actual risk exposure and thus what is inside and outside of scope for analysis. Safety Performance of an AV is defined in terms of:

- Operational Design Domain (ODD) – Broadly, the environment and situations the AV is designed to safely operate within
- Operational Behaviour – The capability of the AV to perform certain actions or manoeuvres or demonstrate behaviours
- Behavioural Competencies – How 'well' the AV performs an action, manoeuvre, or other behaviour

As such, thresholds that delineate between acceptably safe and unacceptably safe performance will also relate to these factors. Other factors may also affect threshold selection such as use case (passenger or goods service). It is therefore necessary to set thresholds in the context of the deployment. There are two ways to accomplish this:

- **Allowing the manufacturer to set acceptable thresholds**. Thresholds would be set in relation to their specific ODD, operational behaviour and behavioural competencies which would be evidenced during approval (i.e. scenario-based testing). This would allow for non-nominal behaviour to be identified very specifically for each AV deployment and so define its own level of acceptable risk. Ultimately this would still need to conform to all legal requirements within the ODD. However, doing so would mean threshold values between manufacturers (or deployments) would differ which inhibits comparability when data is aggregated across all deployments to generate leanings for the industry as a whole.
- **Target thresholds set by the regulator**. The in-use regulator would define acceptably safe performance thresholds that apply to all AV deployments that are reasonably similar. Both leading and lagging SPIs (and the data supporting them) is designed to be outcome based and so technology agnostic. In order to define such thresholds, the regulator would need to provide a set of reference ODDs and applicable behavioural competencies that the manufacturer can match to (as close as possibly). For each reference ODD, acceptable thresholds could be defined. However, given the many variables that make up an ODD, a nuanced change in any of them may affect risk exposure dramatically. Since the correlation between ODD variables and risk is not yet well understood, it may be difficult to generate reference ODDs that allow meaningful comparison.

The choice of which approach will ultimately depend on the priorities of aggregated data analysis for outcome reporting. If generation of industry wide safety recommendations is prioritised, thresholds set by the regulator may be preferred, though the limitations of this approach would need to be clearly stated in any publicly reported data. However, if this is not a priority, then using manufacturer-defined thresholds would allow for better monitoring of safety performance of an individual AV deployment.

A hybrid approach could be considered, whereby initially thresholds are set by each manufacturer. As larger samples of data are collected, greater understanding of the relationship between ODD and risk exposure may be gained and a set(s) of unified thresholds could then be developed in the longer term.

The thresholds for monitoring would likely be governed by the thresholds set as part of the AVs behavioural competencies. As such, the threshold set identifies where there is unsafe

`

performance. However, ultimately, the threshold is set by the manufacturer or regulator, the regulator would need to be prepared to accept or reject the level of risk to the public associated with AV performance in line with those thresholds. In order to make this decision, the regulator would require detailed evidence base for justifying how the threshold value has been set and why it is appropriate.

4.3 Developing correlation through monitoring

One of the key limitations with leading SPIs is that their correlation with actual safety risk is not well known. This relates to the actual measure as well as the threshold values for triggering data capture. It is expected that the value of collecting SPIs will need to be assessed regularly in order to justify their continued recording and specific analysis is needed to confirm the potential degree of correlation. Previous work in this work package has shown a method for this (Chapman and Perren, 2021).

For example, an analysis could show that 10% of the triggers of a leading measure[7] with a specified threshold occurred in an actualised risk outcome (e.g, a collision) whilst 60% occur during clear risk scenarios (e.g., near-collision) and 30% as false positives. This would show that the measure is correlated to both a 10% realised risk + 60% clear risk scenarios which develops the correlation between risk outcomes and potential risk for that measure. This can then be used to justify that the leading measure has 7 times the benefit for statistical analysis than just measuring actualised risk outcomes.

After capture data it should be highlighted that analysis of outputs is essential: imagine instead an alternative trigger again with 10% of cases matching actualised risk, but following analysis of the remaining 90%, 0% can be identified as potential risk scenarios. In this case no potential predictive uplift is possible as no additional risk scenarios are identified. This would present as an appropriate leading measure with no additional correlation to aid risk estimation. These principles apply not just to the measures themselves but also the thresholds selected for them.

The above two scenarios are demonstrated in Table 4 below.

Table 4: Examples assessing predictive value of SPIs

Leading measure example	Actualised risk correlation	Unknown risk requiring sample analysis		Additional Predictive Value
1	10%	60% additional risk scenarios captured	30% false positives	GOOD (*7)
2	10%	0% additional risk scenarios captured	90% false positives	BAD (*1)

[7] Note measures is used over SPI here as this refers to the detection of induvial event

5 Data Analysis Requirements

Metrics may be used to identify individual events for case study analysis. Case studies are a powerful tool to generate in-depth understanding. This is especially beneficial for understanding the causal factors that led to (or precipitated) an event. Case study analysis in an in-use monitoring context relates to in-depth investigation and study of a single event. The scope, requirements, and methods of in-depth investigations for case study analyses has been addressed separately (Arnold and Perren, 2022) and so is out of scope for this document.

In order to evaluate patterns or trends in safety performance, aggregated data analysis over a time period (around 6-12 months is expected to be a reasonable period, although this will vary based on factors such as ODD and distance driven) is then required. Aggregated data is the collection of data from a large number of individual events such as a single collision or near collision event. This aggregated data can be presented in the form of a single, measurable output (such as number of collisions in the last million miles) which represents the vehicle's performance over that period. This output will change with time, as new data is recorded. . Individual data may be collected from multiple sources such as in vehicle data, police reports, investigation data. This data must be combined for the purposes of examining trends and reporting. The main purposes of aggregated data analysis are to examine trends, make comparisons between AV performance and a baseline (e.g. conventional driving performance), or reveal insights that are not observable from a single data point.

For aggregated data analysis to be effective, the results must be normalised by exposure, categorised by variables associated with risk (i.e. segmented). To ensure comparability for datasets of conventional driving, thresholds for the measures need to be defined. The requirements for aggregated data analysis and related issues are discussed further in this section.

5.1 Method of Dataset Generation

For the measures identified in Section 4.1 above, data may be aggregated in two different ways:

- **Rate of occurrence –** where the frequency of SPIs is measured by discrete occurrences below a defined threshold and normalised by exposure (see Section 5.4). This is useful where risk exposure is related to a transitory hazard such as a near miss or a rule violation.
- **Relative duration –** where the time spent below the defined threshold for an SPI is measured as a percentage of overall travel time. This is useful for where risk exposure is more closely related to a continuous hazardous state over a period such as unsafe speeds, close proximity, etc.

Each method applies differently to each measure. Table 5 below assesses the applicability of both methods to each measure. Where both methods are possible it is recommended that both methods are calculated as the presentation of the data in different ways could reveal different insights (for example a single occurrence of TTC could be 1% of the duration of the AVs operation or 20%, but the risk between these is vastly different).

Table 5: Aggregation method relevant to each SPI

	Method of Dataset Generation	
Lagging SPI	**Rate of occurrence**	**Relative Duration**
Collisions	**X**	
Passenger injury	**X**	
Other realised hazards	**X**	
Leading SPI		
Proximity data (TTC, safety envelope, PET)	**X**	**X**
Disengagements (incl. context)	**X**	
Vehicle kinematic data (speed, acceleration, jerk)	**X**	**X**
rule violations	**X**	

5.2 Data Sources

In the proposed in-use monitoring framework, measures are expected to be derived (or at least informed) by a variety of different data sources. The primary source of data is expected to be data collected from the ADS and vehicle systems on board and this will be supported by other data. A summary of the expected data sources for ADS safety performance analysis is given in Figure 2 below.

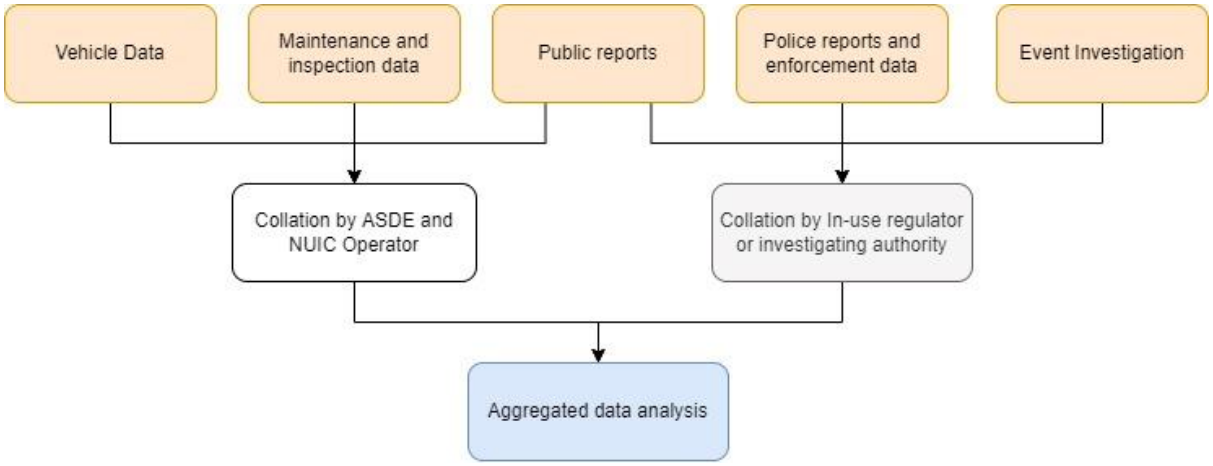

Figure 2: Summary of expected data source for aggregated data analysis

The relevant data sources will depend on the measure type and particulars of the event and will likely be used in combination. For example, a public report of a traffic infraction could be sent to the manufacturer/ operator (directly or indirectly via the regulator) which would prompt the manufacturer to look back at any data recorded at the time of the event to identify relevant details such as location, event partners, etc. (see Section 5.3). This data may also be available from the public report, and potentially other sources.

This leads to the issue of handling overlapping data from multiple sources for analysis. Manufacturers reporting data from disparate sources will need a process to resolve any

conflicts and select the best available data for analysis. This process may be defined by each manufacturer; however, this may lead to different conclusions around the prioritisation of different data sources over others which may hinder the comparability.

5.2.1 Recommendations

It is recommended that further guidance is provided on how overlapping datasets are resolved when reporting aggregated data, providing a robust process with a focus on ensuring the highest quality data is reported. The UK government has set out a Data Quality Framework (DQF) (DQH, 2020) which outlines their philosophy for the management of data quality. It adopts the 6 dimensions of data quality defined by the Data Management Association of the UK (DAMA UK):

1. Accuracy
2. Completeness
3. Uniqueness
4. Consistency
5. Timeliness
6. Validity

The principles defined in the DQF are useful to apply for this context, and guidance should align with assessing and prioritising data based on the 6 dimensions of data quality.

The use of different data sources also requires the recorded parameters to be consistent and translatable for the purposes of comparison. As such the definition of the parameters must be unified across the different data sources. The difficulty in this is the wide variety in the methods of data collection for some measures and parameters. A suitable engineering definition for rainfall for example, may be based on visibility measured by the vehicle's sensors or a record of average rainfall intensity (in mm). These definitions will not be practicable for a public report or for collision investigation. This highlights the need for consistent high-level definitions that are applicable across all data sources which data source-specific definitions can translate into. The in-use monitoring taxonomy (Reed, 2022)defined within this project provides a basis for this.

It is recommended that this taxonomy is progressed and published as part of the in-use monitoring framework. Communication and dissemination with all stakeholders is advised including the police, members of the public, investigating bodies, manufacturers and operators. Where possible, data should align with current processes (e.g. STATS19) so as not to cause undue burden for those collecting and sharing data, especially if voluntary (such as public reporting). The same definitions should also be adopted by or aligned to data sources used for establishing a comparison for conventional driving safety performance. Section 5.3 below considers the different parameters used for analysis (i.e. data segmentation) and proposes potential parameter values based on existing or proposed definitions.

`

5.3 Data Segmentation

Segmentation refers to partitioning data into discrete categories that align with different types of risk. This allows trends and insights to be revealed about safety performance in relation to specific risk scenarios. Many factors may be used for segmentation, but all should ultimately relate to parameters that are expected to impact risk exposure. While there is insufficient deployment mileage to determine parameters that directly correlate, learnings from trials and testing as well as conventional driving safety data can help to identify initial data segmentation parameters to be used that can be correlated to risk exposure as data is collected. Where possible, the data methods for segmentation have been specified such that they align with existing datasets (such as STATS19) to enable utilisation of the data.

A set of segmentation parameters based on perceived factors for risk exposure are proposed below. They are evaluated based on potential values for analysis as well as availability of required data sources and recommendations are made as to how they can be utilised.

5.3.1 Event Classification and Triggered Metrics

A reportable event may be identified through the triggering of one, or a combination of, leading or lagging measure(s) which are monitored throughout operation. At a minimum these measures should include those specified in the proposed minimum dataset (Chapman and Perren, 2021). Furthermore, each event will need to be classified into different types of risky driving or unsafe behavioural competencies.

The type of event (classification) and the measures through which it was identified should be reported.

Value for Analysis:

Matching of leading measures against event type will help to establish their correlation to safety. For example, if low TTC is commonly measured directly preceding a collision, this will identify that low TTC is a suitable predictive measure. Similarly identifying where a threshold triggered but no associated risk event (i.e., a false positive) was identified will also help to evaluate the predictive value of the measure.

Measuring the co-incidence of the measures may also provide insight into trends or patterns of behaviour. For example, if all non-conflict critical incidents involving ODD exits are also measured to have high rates of lateral or longitudinal jerk, this may indicate an issue with the AVs ability to safely perform an MRM, which may prompt further monitoring and investigation.

Data Requirements:

`

Detected events are expected to be processed by the manufacturer in order to establish whether identified events indicated non-compliance with type approval. Through this process, events are expected to be confirmed and categorised into different event types. This means data on the number of false positives and confirmed events as well as an event classification is expected to be available through the proposed in-use monitoring dataset and process.

Recommend that manufacturers report the rate of occurrence of SPIs by event classification in line with event taxonomy definitions.

Recommend that manufacturers report on the metrics used to identify events where are combination of measures are reported.

Recommended Variables:

Event classification is expected to conform to the taxonomy definitions of different event types to provide a standardised mean of event categorisation.

Collision 1: Non-police-reportable low-*g* physical contact	Near-collision
Collision 2: Non-police-reportable property damage only	Safety Critical Event
Collision 3: Police-reportable collision with vehicle / property damage only	Proximity Conflict
Collision 4: Police-reportable collision with possible or slight human injury	Non-conflict critical incident
Collision 5: Police-reportable collision with serious human injury or fatality	Safety relevant violations
False positive (no event)	Road rule violations (may be further sub-divided by infraction type)

N.B: 'Police reportable' refers to where a person(s) is injured or if there is an immediate risk of injury or death (Arnold and Perren, 2022)

5.3.2 Event partners

HAVs are expected to interact with many different road users during operation. Correct identification of an object (including road users) and planning an appropriate action is led by the ADS' Object and Event Detection and Response (OEDR) capability. It is important to understand whether the ADS can safely manage interactions with all road users and identify where there are road user demographics that are disproportionately affected by the AV.

Value for Analysis:

Measuring the rate of occurrence of unsafe events by the party involved in the event can help to identify which road user groups and demographics are most exposed to the risk of potential unsafe driving by the AV. This may give insights into the performance of Object and Event Detection and Response (OEDR) capability, notably perception and classification of different

objects. Identifying whether there is disproportionate risk to road user demographics from AVs is a key recommendation from the Law Commissions' and a likely public expectation.

Monitoring the demographic of involved parties may also identify biases in machine learning training sets used for the perception and classification of road users or defects in ODD construction so that they can be corrected. Algorithmic decision making such as Machine Learning can introduce bias towards different demographics. This has been shown to inadvertently discriminate against marginalised populations (Mittelstadt *et al.*, 2016). In the context of automated vehicles this may unintentionally introduce disproportionate levels of risk to road users with protected characteristics through for example, the bias in path prediction of pedestrians with different skin tones. While the exact nature of how algorithmic bias may impact AV safety is not known, it is recommended that it is monitored (CDEI, 2020). In order to monitor against this, data around protected characteristics of event partners would need to be collected.

Data Requirements:

It is possible to collect data on the category of event partner directly from vehicle data if the AV is capable of identifying and classifying objects in such a way. As such, data could be retrieved on object classification and matched to the event in question; however this may not always assign the correct object as the "event partner" accurately and there is also the possibility of objects being classified incorrectly. In these cases external sources of data (such as public reported information or infrastructure data), or video data would be needed to support this classification.

Furthermore, ADS' do not need to explicitly classify objects in order to drive safely and comply with traffic rules and it is known that some ADS solutions do not have this capability. In such cases, post-event processing and analysis would be required to identify and classify the event partner after the fact, which may require the matching of additional datasets such as video data. These are outside the minimum recommended dataset for in-use monitoring proposed earlier in the WP (Chapman and Perren, 2021).

Collection of data relating to protected characteristics (as defined by the Equality Act) is controlled by the Data Protection Act and most are treated as special categories of personal data. While restrictions on the collection, storage and use of special category data applies, the DPA explicitly allows it for the purposes of monitoring equality (CDEI, 2020). In fact, the Centre for Data Ethics and Innovation (CDEI) suggests that a greater collection of protected characteristic data would allow for fairer algorithmic decision-making in many circumstances and is recommended (CDEI, 2020). This could be extended to consider factors such as whether the vehicle accurately classifies wheelchair or mobility scooter users as pedestrians. Practically the collection of such data by the vehicle in real-time may be incredibly difficult to do consistently.

For casualties, some protected characteristics such as, age and race are recorded on STATS19, however age may be estimated by the recording officer rather than established as fact. Ethnicity is recorded for all casualties (including fatal, serious, and slight). Officers will ask for self-defined ethnicity using the 16+1 standard ethnicity categories (MPA, 2007). Where the casualty cannot self-define the visible 6+1 categorisation is used by the officer. TRL experience

with police investigation has outlined, however, that officers do not often record ethnicity because it rarely pertains to event causation. As such, this data may be unreliable.

Recommend that the rate of occurrence of events is categorised by event partner (including protected characteristics) for all collision events where an investigation takes place.

Recommend that event partners are identified and classified for all reported events. Where possible, collection of data relating to some or all protected characteristics for all events is highly encouraged.

Recommended Variables:

Event partners are expected to be reported in line with the in-use monitoring taxonomy (Reed, 2022) which aligns with categories within the STATS19 form for police reported collisions in Great Britain:

Car	Motorcycle – cc unknown
Taxi / Private hire car	Electric motorcycle
Van ≤3.5t mgw	Pedal cycle
Goods vehicle 3.5t<mgw<7.5t	Bus or coach ≥17 passenger seats
Goods vehicle ≥7.5t	Minibus 8-16 passenger seats
Goods vehicle – weight unknown	Agricultural vehicle
Motorcycle ≤50cc	Ridden horse
Motorcycle >50cc and ≤125cc	Mobility scooter
Motorcycle >125cc and ≤500cc	Tram / Light rail
Motorcycle >500cc	Other
Micromobility User	

Protected Characteristics are defined in the Equality Act. Some are less likely than others to be subject to algorithmic bias than others. A subset of the protected characteristics likely to be of interest is proposed below:

Sex	Age
Race, colour, ethnic, national origin, nationality	Pregnancy and maternity
Disability status	

5.3.3 Weather and Environmental Characteristics

Environmental factors such as weather are widely reported to impact an automated system's OEDR capability as well as the dynamics of the vehicle (i.e., loss of grip/traction)). Safety performance will depend on the AVs ability to correctly perceive objects and the situation. Sensor performance in adverse weather conditions will likely have an impact on safety.

Value for Analysis:

Reporting on weather and environmental characteristics can help identify operational limitations in regard to weather, which can inform policy on restrictions in certain weather conditions.

Data Requirements:

Weather and environmental data is not directly specified in the minimum dataset. After-the-fact analysis may allow this data to be collected. Some LiDAR based approaches have been used for vehicle detection of rain and fog (Yoneda *et al.*, 2019) which can be used for ODD monitoring, but it is not easy to provide in a consistent format. This may also require a large degree of post-processing to retrieve an accurate measurement and will place additional burden on developers to store potentially large volumes of raw LiDAR data.

Alternatively, it may be possible to match event time and location with meteorological data through analysis post-event. However, rain and especially fog can be highly localised, in time and space. Standard meteorological data is likely not specific enough to draw conclusions about conditions at the event location. Such a process is also likely to be human led and resource intensive.

At a minimum it is expected that the ADS will be aware of conditions outside of its ODD and be able to initiate and safely complete a Minimum Risk Manoeuvre (MRM). As such disengagements and MRM events can be reported and relatively little investigation should identify whether the cause of the disengagement is related to limited sensor performance, which may imply adverse weather impacts. For example, the AV may detect reduced visibility outside a limit defined in its ODD, while the AV may not classify it as rain or fog, or an obscured sensor it would imply that environmental conditions are at play. Further classification of weather at the time of the event will be extremely useful and doing so is strongly encouraged, however, the burden on the manufacturer is currently unknown.

BSI PAS1883 specifies a method for recording environmental conditions for specifying and ODD. However, the methods and measurements proposed are not likely to be measurable by a vehicle (such as rainfall measured in mm/h) or a person attending the scene or reviewing the data (BSI, 2020).

`

It is possible to classify weather and environmental characteristics qualitatively from on-site attendance of event (applicable where a collision investigation takes place). This is done for all police reported collisions currently (HM Government, 2021).

Recommend that weather and environmental characteristics are reported alongside rate of occurrence for all collision events where an investigation takes place. This is expected to be available from collision investigation.

Recommend that disengagement/MRM status are reported alongside rate of occurrence for all reportable events. Where necessary the reason for disengagement/MRM should be investigated and where weather and environmental characteristics are the cause, this should be reported.

Recommended Variables:

Weather and environmental characteristics align with classification used at the scene by Police investigators in STATS19:

Fine (without high winds)	Fine (with high winds)
Raining (without high winds)	Raining (with high winds)
Snowing (without high winds)	Snowing (with high winds)
Fog or mist	Other
	Unknown

Light conditions should also be reported:

Daylight	Darkness: no street lighting
Darkness: streetlights present and lit	Darkness: streetlighting unknown
Darkness: streetlights present but unlit	

5.3.4 Static Operational Domain Elements

Value for Analysis:

Collection of road environment data may help to identify trends or correlations between rate of occurrence of unsafe events and specific road features. This may indicate a weakness in ODD coverage of the Target (i.e. deployment) Operational Domain (TOD) which may then prompt further investigation and resolution by the manufacturer. It may also reveal insufficiencies in the scenario test programme for certain road features orstatic objects and may also indicate areas where road design can be improved to accommodate AV deployments in the future.

Data Requirements:

For the purposes of the proposed in-use monitoring scheme, the manufacturer is required to share the world model representation the AV perceives and is used as an input to its planning module[8]. The model should include the objects that are required for the AV to be able to demonstrate its defined behaviours and competencies and ensure compliance with the

[8] This concept is discussed further in the In-Use monitoring framework report for this project (Balcombe and Perren, 2022).

`

Highway Code. In order to ensure compliance with the Highway Code, the world model should contain the static objects referenced in the Highway Code. Examples include lane markings, stop line markings, pedestrian crossings, traffic lights (including status). As a result, the static objects perceived at the event by the vehicle should be reportable. However, this does not provide insight for events where objects were not classified or misclassified objects which could be an important factor in event causation.

Furthermore, there is no standardised or prescribed list of objects that must be detected and classified by an ADS. The need to identify and classify different objects to ensure rule compliance may also differ between systems. For example, an AV may not be able to identify a stop sign but may demonstrate compliant behaviour by assuming a stop sign exists at every junction and thus broadly complying with the relevant highway code rules (though this may also introduce risk as it is counterintuitive to human behaviour). In order for this data to be shared, a reference list of relevant world model objects could be provided as guidance to manufacturers. This would need to be non-prescriptive to allow for different technical solutions but should be broadly based on the objects listed within the highway code.

After-the-fact analysis of the event may also allow road features and other static objects to be identified such as through matching location to map data, although real time processing of the data onboard the vehicle is expected to be much less resource intensive.

It is possible to classify road type and road features qualitatively from event attendance and investigation as is done for all police reported collisions currently (HM Government, 2021). STATS19 forms provide a coding method for this.

`

Recommend that static operational domain elements that are relevant to the event in question are reported alongside rate of occurrence of event. To achieve this, a reference list of relevant objects should be produced that can be used to ensure relevant objects from the world model are published as reportable outputs.

Recommend road features and other static objects are recorded and reported for all investigated collisions as is done currently. In order to align statistics for collision reporting and other events (where no on-site investigation takes place), it is recommended that the definitions of road features and static objects are aligned with existing collision reporting methods (i.e. STATS19).

Recommended Variables:

Road features and other static objects should align with the classification used at the scene by Police investigators in STATS19. The following factors should be reported for all investigated collisions. Alignment should be sought between these, and elements shared by the world model:

Road type (roundabout, one-way street, dual carriageway, single carriageway, slip road)	Pedestrian crossing (zebra crossing, pelican/puffin/toucan, traffic signal (pedestrian phased), footbridge, subway, no physical crossing within 50 m)
Junction detail (roundabout, mini roundabout within 20m of junction, slip road, T junction, Private driveway/ entrance)	Road surface Condition (dry, wet/ damp, snow, frost/ ice, flood)
Junction control type (authorised person, automatic traffic signal, stop sign, give way/ uncontrolled)	Carriageway hazards (dislodged vehicle load, other object, involvement in previous accident, pedestrian in carriageway, animal in carriageway)
Special conditions (auto traffic signal out, auto traffic signal partially defective, permanent road marking/ signing defective or obscured, road	

5.3.5 Event Type

For many decades, TRL has attended road collisions to gather additional data that can help to understand why and how incidents occurred and to help develop possible countermeasures to prevent their occurrence in future. The Road Accident In-Depth Studies project (RAIDS; TRL, 2012), which TRL manages on behalf of the Department for Transport, uses a specific coding system to characterise the most common types of incidents. This system uses fifteen different manoeuvre types (e.g. overtaking and lane change; collision with obstruction; merging) each with between one to seven different variants (e.g. cornering – lost control cornering right; cornering – lost control cornering left; cornering missed intersection or end of road).

Value for Analysis:

A similar categorisation of manoeuvres for safety events involving LSAVs may be helpful for regulators in determining whether it is safe for operations to continue and for developers and operators in taking mitigating actions to prevent such events happening in future. This may also help to identify trends in high-risk manoeuvres for all AVs or issues with aspects of an individual AV's behavioural competencies.

Data Requirements:

Manoeuvre planning is likely to be a key element in understanding event causation and as such, the manufacturer would be expected to make this data available to provide context to the event. This data is expected to be available, but the ontology and classification of manoeuvres done by an AV would likely be specific to each ADS' planning module and would not be consistent across different AVs.

Recommend that HAVs report event type. This should be supported by standardised, objective method by which vehicle data could be used to report event type automatically when an event has been detected.

Recommended Variables:

RAIDS collision codes provide a useful starting point for classification, but their definitions would likely need adapting to account for non-collision scenarios that are also proposed to be reported on under the scheme. Codes are listed below, but for each code, manoeuvre/collision sub-types exist.

Overtaking and lane change	Crossing (vehicle turning)
Head on	Merging
Lost control (straight road)	Right turn against traffic
Cornering	Manoeuvring
Collision with obstruction	Pedestrians (crossing road)
Rear end	Pedestrians (other)
Turning vs same direction	Misc.
Crossing (no turns)	

5.3.6 Causal and Contributory Factors

It is useful to clarify why each event occurred to help understand the chain of accountability for an event and to consider how such events might be prevented in future. Determining responsibility for causing an event can be challenging with fault potentially lying across multiple actors. However, in the deployment of LSAVs it will be vital to establish whether an incident was caused or influenced by operation of the vehicle or whether responsibility lay elsewhere.

Value for Analysis:

Analysing the rate of occurrence of events by suspected cause will help to identify insights into common issues and failures experienced by AVs both individually (to feedback to the manufacturer for improvement) and for all AVs deployed under the scheme (to develop industry wide recommendations.

Data Requirements:

It is expected that upon identification of an event, the manufacturer should record and persist all data necessary to determine the cause of the incident. This can then be accessed by the regulator for investigation or event causation is self-reported by the manufacturer. The degree to which this data is used to investigate to determine causation is likely to be a decision made by the regulator (although an manufacturer may wish to proactively investigate events prior to reporting) depending on the type of event.

One of the issues with this approach, however, is the ability to correctly understand all the causal factors associated with the event. Learning from the use of contributory factors in police reporting via STATS19, is that it is not always easy to establish all the causal and contributory factors involved in an event from initial investigation. There is often a bias towards factors that are readily apparent, or more ubiquitous. For example, TRL experience has found "failed to look properly" and "failed to judge other person's path or speed" are often reported as contributory factors, because they apply to almost all possible collisions scenarios. There are concerns as to the ability of manufacturers to be able to establish event causation reliably and consistently. However, it should be said that this limitation does not necessarily prohibit reporting of it. More detailed investigation may be used to more accurately understand and report of causal and contributory factors for a subset of events as well as a means of determining how accurately they are assessed during initial investigation.

Recommend event causation is reported for each event for high-level causal and contributory factor definitions. These may be at first defined by the manufacturer. For collisions or other events investigated by the regulator, this may be reaffirmed or updated following the results from the investigation.

Recommend the regulator regularly assesses the applicability of causal and contributory factors and how well factors reported by a manufacturer match those determined through further investigation.

Recommended Variables:

The following event causal factors have been defined in the In-use monitoring taxonomy. Detailed event investigation may identify new types or subsets of factors, and so should be used to update these variables over time. (Reed, 2022)

Perception error	Infrastructure / Environment
Decision error	Other Road User action/error
Action error	Cybersecurity
Human factors error	Outside ODD

5.4 Normalisation by Risk Exposure

Normalising results refers to adjusting data with different measurement scales and risk levels to enable comparison across data sets and correct for known limitations (AVSC, 2021). Early AV deployments will accumulate far fewer miles than conventional vehicles in the same time period. As such, the count of risk events cannot be directly compared between ADS and conventional driving. Exposure normalisation improves the comparability of results across different data sets by creating a common scale and improves interpretability and portability of results.

To enable in-use monitoring assessment of safety performance, data should be adjusted by a measure of risk exposure. Commonly traffic collision statistics are normalised by number of kilometres driven or hours of operation (AVSC, 2021). However, a meaningfully large sample size cannot be obtained for early deployments using these measures.

The Automated Vehicle Safety Consortium (AVSC) suggests utilising other exposure factors such as frequency of road user interactions or frequency of scenario types, can help to increase sample size for limited deployments and also enable comparison between deployments with different ODDs (AVSC, 2021). Consider the rate of occurrence of traffic infractions, for example ignoring a red traffic light signal. While kilometres travelled in a single journey may be low, the number of signal-controlled junctions travelled may be much higher and so its use as a measure of risk exposure would generate a much larger dataset more quickly and be much more specific to the driving context.

Allowing different sampling techniques to be used by different manufacturers would make the comparison and aggregation of industry-level data impossible. Furthermore, a sampling strategy may be selected that favours a particular ODD or use case which may skew evaluation of safety performance. Ensuring consistent methods for sampling and presentation of data is essential for public comprehension. As such it is advised that that the regulator specifies a sampling strategy for which defines that an acceptable measure of risk exposure for each defined measure. Care must be taken, however, to ensure that the data required for the measures chosen can be collected and do not introduce any technological bias.

5.5 Comparison against Conventional driving

The Law Commission of England and Wales and the Scottish Law Commission, in their joint report, discuss the subject of automated vehicle (AV) safety and how the Safety Standard for allowing AVs on the road should be set. During the Law Commissions' consultation process, the majority of consultees who did not express their preference for "other", expressed their preference for a Safety Standard that would require AVs to be safer than the average human driver. With every option in the consultation referencing the performance of a human driver, this automatically highlights the requirement to measure the safety performance of AVs and compare it to that of human drivers.

Safety performance for conventional driving is commonly measured using safety statistics such as collision rates, injury severity and various other indicators generated through data collected post-collision (usually through police reporting or collision investigations). For example, a key safety indicator is the number of fatalities per billion miles travelled. In Great Britain, the fatality rate per billion miles travelled was 4.87 in 2019 (pre-covid levels). For AVs to demonstrate improved safety performance with regards to severe collisions that lead to fatalities, the fatality rate would need to be lower, in a statistically significant way.

Safety statistics are generated via the processing of datasets contained in relevant databases such as the STATS19 and RAIDS databases. The difficulty in comparing conventional and AV driving through safety statistics such as the above lies with collecting the required data to build a sample large enough to make a meaningful comparison; the low occurrence of collisions with fatalities/injuries per miles travelled, combined with the small number of AVs on the road would make this data collection effort an onerous task (see Section 4.1). It should also be noted that human driving performance changes slightly on a yearly basis (gradual reduction in fatalities and injuries, partly due to improvements in vehicle safety technology) which would make meeting human performance standards a moving target.

`

Collection of data through different sources and generation of both lagging and leading SPIs is the only feasible solution to provide an acceptable safety standard in terms of conventional driving safety performance. The main source of data will inevitably be the vehicle itself (in-vehicle monitoring data from the ADS log, EDR and DSSAD devices, and Naturalistic Driving Studies data for conventional vehicles, AVs collecting data as part of their operation) with other sources supporting (police report data, infrastructure data, maintenance data, public reports, and collision investigation data – see Section 5.2 for additional detail). It should be considered when building a dataset for comparing conventional and AV driving that, any risk scenario and performance evaluation will require a data segmentation process according to relevant factors (event classification, event partners, weather and environmental conditions, road type, event type and contributory factors – see Section 5.3 for additional detail). This is a prerequisite for comparing driving performance in specific scenarios and will also affect the required data sample for each case.

A set of key lagging and leading SPIs with a clear correlation to safety is presented in Section 4 of this report, with many of the recommended SPIs valuable in assessing safety in both conventional and AV driving. There is, however, a number of SPIs that are not appropriate for a comparison of this type, as they are not relevant to conventional driving. For lagging SPIs, those would be MRM activation, ODD exit, autonomous sensor fault triggers and any others that are linked specifically to automated operation. Similarly, there are leading SPIs that would not be of any use in a conventional/AV driving comparison (for example roadcraft, disengagements, traffic infractions or other SPIs not easily quantifiable).

A comparison between conventional and AV driving is feasible when using measures that can be defined quantitively for both automated and conventional driving populations, such as TTC, acceleration or using proximity criteria.

5.5.1 Recommendation

It is recommended that the specific research activities are required to generate an initial data baseline for human driving of the performance measures, considering the need for datasets that are comparable with AV data, segmented by the same risk exposure factors, and represent scenarios and risk experienced during the AV deployment. In time, data collection may be possible through conventional driving data recorders such as EDR, however the dataset would need to align with that required for In-use monitoring to enable comparability. In order to establish a scheme initially, A naturalistic driving study may be a useful starting point to give the required data to set an initial standard.

`

6 Summary and Recommendations

The joint Law Commissions report on automated vehicles outlined numerous methods for regulators to assess and report safety performance in line with a defined safety standard; these options are discussed in Section 3. Their key recommendation in this area was that Ministers should set an appropriate safety standard for automated vehicles with support from experts (Law Commission and Scottish Law Commission, 2022). No matter how the safety standard is developed, the in-use regulator will be expected to develop practical ways of measuring and reporting on current safety performance against the standard.

The method by which safety performance will be reported is therefore highly dependent on how the safety standard is formulated. There are broadly two aspects that are proposed:

- Assessment against a defined standard (linked to some level of safety performance); and
- Demanding a higher standard over time with the aim of gradually improving safety performance.

In this report, we suggest that a single measure of performance based on traffic safety statistics is not sufficient. Consultees of the Law Commissions' work also highlighted that it has never been feasible to provide an collision rate comparison between a new transport system and its predecessor (Law Commission and Scottish Law Commission, 2022). Rather, a combination of both leading and lagging measures of safety performance are required that monitor different elements of safety. This report evaluates a series of SPIs (both leading and lagging) that may be used to evaluate safety performance, derived largely from current methods of tracking safety performance.

Collection of leading measures allows for much larger datasets to be generated compared to the data available from solely measuring lagging measures such as collision rate. However, while leading SPIs are a crucial element they have their drawbacks. The key concern is to what extent the SPIs correlate to an increased exposure to risk relating to AVs, and so to what extent are they useful to monitor. This cannot be known prior to deployment. As such, this report has taken a qualitative approach to evaluate the usefulness, practicablity, and perceived safety correlation to prioritise a set of leading measures that will allow useful data to start being gathered. This leads to a key finding: while collecting, reporting and evlauating the data may be burdensome, and some SPIs may be found to be not useful, these matters will never be resolved until data starts to be collected. It is imperative therefore, that some amount of data starts to be collected as soon as possible so that the process for outcome reporting can be refined over time.

The measures and SPIs in this report are therefore a 'best estimate' of the most meaningful methods of monitoring safety performance which serve as a starting point and which we anticipate continually evolving over time as learn more about AV risk factors and how they may be detected, analysed and compared. This report also describes how the effectiveness of the monitoring methodology (including SPIs) can be evaluated during its operation.

However, it is well known, that for the the proposed SPIs to be meaningful in any way, they must account for the variables that affect risk exposure, such that different ODDs, deployment contexts and use cases may require different targets (or target values) to assess

the SPIs against. As such, methods are proposed to segment data by key variables known (to the best of our current knowledge of AVs) to have an impact on risk exposure. These variables are event classification and type, event partners, weather and environmental charcteristics, road type/ODD elements, and the causal and contributory factors associated with the event[9]. For each of these variables, the value of capturing them is assessed alongside the availability of the data required to capture them effectively. For each, recommendations are made as to what variables are recommended, how the relevant data is captured and what values (or categories) they may take.

Early AV deployments will accumulate far fewer miles than conventional vehicles in the same time period. As such, the rate of occurrence of risk events cannot be directly compared between ADS and conventional driving. Methods for normalising data sets are required that factors in the exposure to risk. It is not recommended that normalisation be based solely on vehicle miles travelled (which is the usual method for traffic safety analysis), as it will take significant time to generate large sample sizes where statistical significance can be drawn. Instead other variables to normalise against are advised. The best method of normalisation will likely relate to the predominant factors that affect risk exposure for a particular AV deployment, as such the normalisation method could vary. However, without a consistent means of normalisation across manufacturers (and their deployment) the regulator will not be able to aggregate data to evaluate safety performance industry wide. Guidance is required to provide manufacturers with a consistent method of how to present Safety performance data which should align where possible with international approaches to establish interoperability and access to much larger datasets in time.

The proposed in-use monitoring scheme primarily relies on the capture of in-vehicle data to monitor AV safety, which is a key input of aggregated data analysis for monitoring safety performance. However, the use of other data sources, such as investigation data, public incident reports and police reports cannot be overlooked as these provide coverage of some safety relevant events that the vehicle may not or can not detect itself. With the aggregation and processing of disparate data sources brings issues of conflicting and duplicating data which affect the quality of the assessment. manufacturers and vehicle operators will need to have a data management plan in place for handling and priortising between data sources. Guidance may be required to ensure this is done consistently.

There is a need to assess safety performance against a defined standard, with a preference for the standard to be set in line the safety of conventional driving. In order to do this a baseline level of safety performance for conventional driving needs to be established. Currently, while traffic safety statistics are collected in GB and eslewhere, there main focus is on police attended collisions and traffic infractions. To compare with the wealth of data collected by an AV, equivalent datasets for leading SPIs and risk variables need to be collected. In practice, this means human driving perfromance will need to baselined in comparable, scenarios, use cases and deployment contexts. A naturalistic driving study with a methodology consistet with the data required in for the in-use monitoring schem would likely be the best way to collect this data.

[9] These have been defined in the Road Incident Taxonomy report for this project (Reed, 2022)

`

The findings of this work have been summarised into a high-level process for aggregated data analysis which is shown in Figure 3. The associated process steps have then been assigned to different stakeholders through the use of a RACI (Responsibility, Accountability, Consulted, Informed) matrix to highlight how each stakeholder is involved in the process. This is shown in Table 6, accompanying the process flow.

`

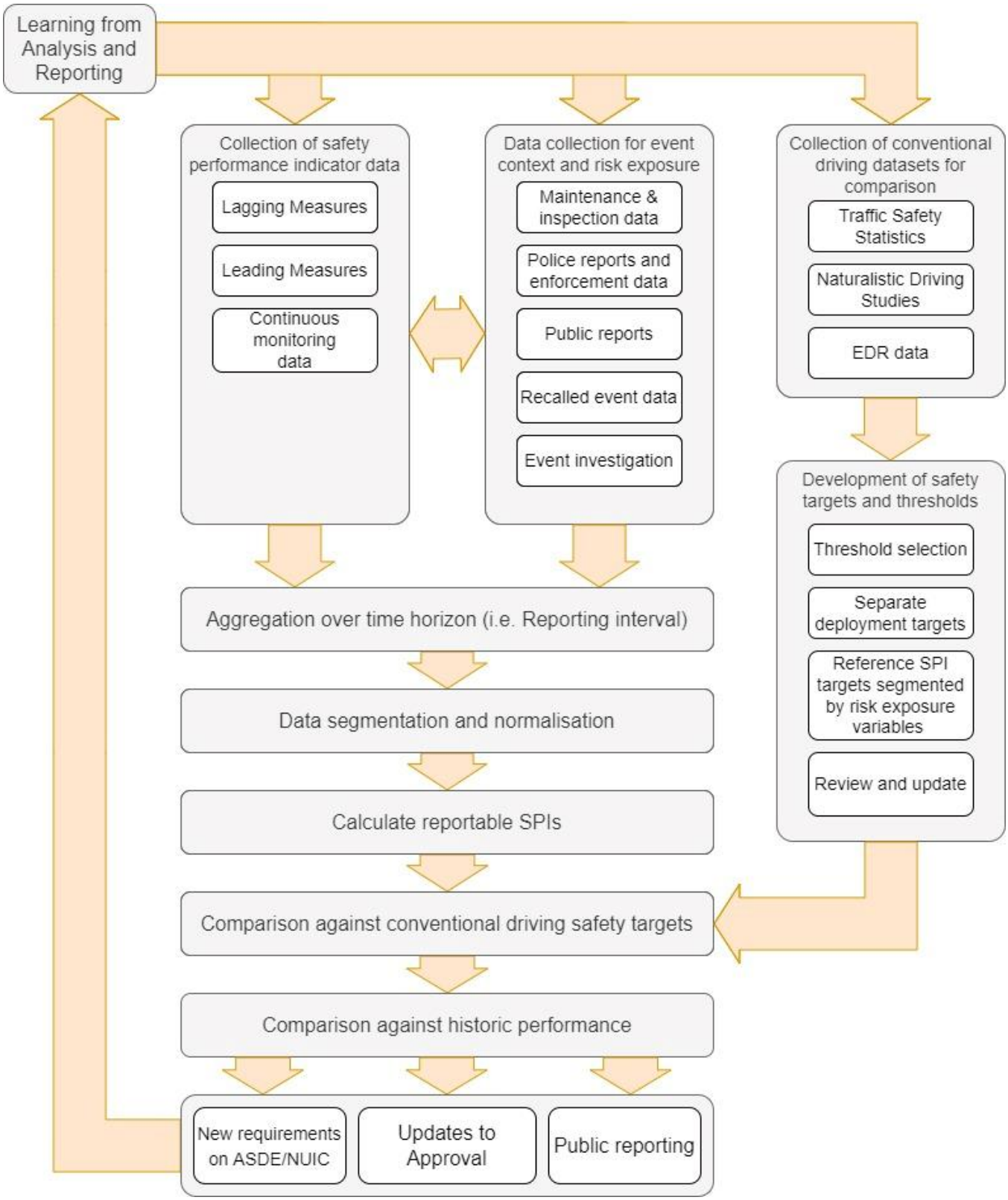

Figure 3: Proposed safety performance reporting process

Table 6: RACI Matrix for the proposed safety performance reporting process

	In-Use Regulator[10]	Approval Authority[11]	Manufacturer	Operator
Collection of safety performance indicator data	I		A	R
Data collection for event context and risk exposure	I		A	R
Aggregation over time horizon (i.e., Reporting interval)	I		A	
Data segmentation and normalisation	I		A/R	C
Calculate reportable SPIs	I		A/R	C
Collection of conventional driving datasets for comparison	A/R	C	C	I
Development of safety targets and thresholds	R	A	C	I
Comparison against conventional driving safety targets[12]	A/R	I	I	I
Comparison against historic performance[13]	A/R	C	I	I
Placing new requirements on manufacturer/operator, Updates to type approval, public reporting	R	A	C	C

The purpose of this reporting process is threefold.

- To feedback on safety performance to the manufacturer, in order to assess compliance with the scheme and set out any necessary remedial action
- To generate learnings around limitations of the approval scheme, that need to be addressed; and

[10] In-use regulator also includes any investigating bodies.

[11] And Authorisation Agency, as appropriate

[12] This comparison would be provided only to the manufacturer it relates to. Scheme-wide statistics would be produced and which are shareable with public and wider industry

[13] This comparison would be provided only to the manufacturer it relates to. Scheme-wide statistics would be produced and which are shareable with public and wider industry

`

- To report on the safety performance of automated vehicles in GB for the purposes of generating industry wide knowledge and assure the public that there is sufficient and robust oversight of these vehicles.

In order to ensure that the reporting continuously meets to needs and expectations of the public, it is recommended that the in-use regulator regularly reports to Transport Focus and the Office of Road and Rail (ORR).

`

7 References

Arnold C and Perren W (2022). *GB Safety and Secturity Approval Scheme: Post event investigation framework*. TRL: Crowthorne, Berkshire.
AVSC (2021). *AVSC Best Practice for Metrics and Methods for Assessing Safety Performance of Automated Driving Systems (ADS) AVSC00006202103*. Society of Automotive Engineers.
Balcombe B and Perren W (2022). *GB Safety and Security Approval Scheme: Safety Monitoring Framework*. TRL: Crowthorne, Berkshire.
BSI (2020). *PAS 1883:2020 Operational Design Domain (ODD) taxonomy for an automated driving system (ADS) – Specifi cation*. British standards Institute.
CDEI (2020). *Review into bias in algorithmic decision-making*. CDEI: London.
Chapman S and Perren W (2021). *In-use monitoring dataset specification (draft)*. TRL.
Cuerden R and McCarthy M (2016). *The methodology and initial findings for the Road Accident In Depth Studies (RAIDS) Programme*. TRL.
DfT (2020). *Reported road casualties in Great Britain: 2019 annual report*. DfT.
DQH (2020). *The Government Data Quality Framework*. HM Government: London.
European Commission (2020). *Ethics of connected and automated vehicles : recommendations on road safety, privacy, fairness, explainability and responsibility*. European Commission, Directorate-General for Research and Innovation, Publications Office: https://data.europa.eu/doi/10.2777/035239.
European Commission (2022). *EC Delegated Regulation C(2022)395*. European Commission: Brussels.
GIDAS 2019, *GIDAS - German In-Depth Accident Study.*, <https://www.gidas.org/start-en.html>.
HM Government 2021, *stats19-forms-and-guidance.*, <https://www.gov.uk/government/publications/stats19-forms-and-guidance>.
Kalra N and Paddock SM (2016). *Driving to Safety: How Many Miles of Driving Would It Take to Demonstrate Autonomous Vehicle Reliability?* RAND Corporation: Santa Monica, CA.
Kyriakidis M, Happee R and de Winter JCF (2015). Public opinion on automated driving: Results of an international questionnaire among 5000 respondents. *Transportation Research Part F: Traffic Psychology and Behaviour, Volume 32*, Pages 127-140.
Law Commission and Scottish Law Commission (2022). *Automated Vehicles - Final report*. HM Government: London.
Mittelstadt BD, Allo P, Taddeo M, Wachter S and Floridi L (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*.
Mobileye (2020). *Implementing the RSS Model on NHTSA Pre-Crash Scenarios*. https://www.mobileye.com/responsibility-sensitivesafety/rss_on_nhtsa.pdf [Retrieved on 7th August 2020].
MPA 2007, *The code systems used within the Metropolitan Police Service (MPS) to formally record ethnicity.*, <http://policeauthority.org/metropolitan/publications/briefings/2007/0703/index.html>.
National Highway Traffic Safety Administration (2006). *The 100-Car Naturalistic Driving Study: Phase II - Results of the 100-Car Field Experiment*. US Department of Transportation: Washington DC.
Nvidia 2020, *PLANNING A SAFER PATH.*, <https://www.nvidia.com/en-gb/self-driving-cars/safety-force-field/>.

`

Penmetsa P, Adanu EK, Wood D, Wang T and Jones SL (2019). Perceptions and expectations of autonomous vehicles – A snapshot of vulnerable road user opinion. *Technological Forecasting and Social Change, Volume 143*, Pages 9-13.
RAND (2018). *Measuring Automated Vehicle Safety - Forging a Framework*. RAND: Santa Monica, California.
Reed N (2022). *GB Safety and Security Approval Scheme: In-use monitoring taxonomy*. TRL: Crowthorne, Berkshire.
Yoneda K, Suganuma N, Ryo and Aldibaja M (2019). Automated driving recognition technologies for adverse weather conditions. *IATSS Research, Volume 43, Issue 4,* Pages 253-262.

`

Automated Vehicle Safety Assurance - In-Use Safety and Security Monitoring

Abstract

This report discusses the potential uses of in-use vehicle data to generate aggregated data to calculate Safety Performance Indicators (SPIs) and track safety performance of Automated Vehicles (AVs) throughout their deployment lifetime. The two primary benefits of collecting this data are to provide a feedback loop to AV Manufacturers and Operators to improve their safety performance as well as compare the safety of AVs more broadly against conventional driving and other transport modes. This work identifies a set of SPIs that can be recorded using in-vehicle data and other available data sets that can form the basis of a monitoring process that can be improved over time. It also discusses how the likely data sets should be processed and analysed to provide fair and accurate comparisons to conventional driving. Based on this work, a high-level process for outcome reporting by an In-Use Regulator has been proposed.

Relevant Reports

- Automated Vehicle Safety Assurance - In-Use Safety and Security Monitoring Task 1 – Road Incident Taxonomy; https://doi.org/10.58446/mvuc1823
- Automated Vehicle Safety Assurance - In-Use Safety and Security Monitoring Task 2 - Minimum Dataset Specification; https://doi.org/10.58446/nksn4732
- Automated Vehicle Safety Assurance - In-Use Safety and Security Monitoring Task 3 - Safety Monitoring Framework; https://doi.org/10.58446/sgxq7004
- Automated Vehicle Safety Assurance - In-Use Safety and Security Monitoring Task 4 - Post Event Investigation Process; https://doi.org/10.58446/egfa6491
- Automated Vehicle Safety Assurance - In-Use Safety and Security Monitoring Task 6 - Data Privacy; https://doi.org/10.58446/dwll8689
- Automated Vehicle Safety Assurance - In-Use Safety and Security Monitoring Task 7 - Change Control; https://doi.org/10.58446/bpdl3309

TRL
Crowthorne House, Nine Mile Ride,
Wokingham, Berkshire, RG40 3GA,
United Kingdom
T: +44 (0) 1344 773131
F: +44 (0) 1344 770356
E: enquiries@trl.co.uk
W: www.trl.co.uk

ISSN: 2514-9652

DOI: https://doi.org/10.58446/qlpq9096

PPR2020