Data Management in AI-Based Safety-Critical Systems

Ian Marsh
Collaborative Autonomous Systems
RISE Research Institutes of Sweden AB
Stockholm, Sweden
ian.marsh@ri.se

Thanh Hai Bui
Mobility and Systems
RISE Research Institutes of Sweden AB
Gothenburg, Sweden
thanh.bui@ri.se

***Abstract*—Few sectors demand stricter safety guidelines, standards and systems than the automotive one. Brands often compete on their ability to promote and produce a safe driving experience. Closely related to safety in physical applications is the issue of trust. A trusted and safe brand is ultimately a sellable one.**

A lower acceptable limit for Automated Driving (AD) vehicles has been stated as *at least as safe as a human driver*. Achieving this puts high requirements on perception. This is non-trivial as humans spend many years learning about environments, i.e. spotting subtle differences and avoiding potential dangers.

Machines might not be able to gather, learn and process data over many years, but more importantly, cannot generalise as humans do, especially with respect to danger. Therefore, with an eye on safety, this paper looks at sensors, impairments, data quality and its impact on downstream machine learning. Poor quality data increases the risk of ML performance degradation which ultimately may cause accidents.

Poor data requires additional cycles of the drive-collate-process-assess process. A complicating factor is adverse weather, which in this work we factor in with respect to Nordic conditions. Whilst frameworks and standards have already been introduced, real use cases are needed to operationalize the standards and introduce useful and relevant concepts, that ensure safe automated driving. Namely, we introduce data readiness levels, akin to TRLs for data, data descriptors and implementation of explainable AI. We also use a data inspection tool for checking all sensor modalities *at the same time*, to 'debug the data'.

***Index Terms*—Automated driving, Data management, Safety-critical, AI trustworthiness.**

I. Introduction

The adoption of AI is a key success factor for safety-critical autonomous systems, particularly given the rapid advancements in the sector. However, integrating AI into safety-critical systems introduces a new set of challenges resulting from uncertainty in AI models. In this paper, uncertainty refers to the *difference* between training and deployment. Specifically, it refers to models trained to navigate a vehicle.

Without an appropriate safety assurance strategy, uncertainties in AI predictions can lead to unexpected hazardous situations that impact human safety [1]. Unlike traditional software, AI models heavily depend on data, which conditions both their performance and robustness. This dependency may lead to unpredicted, and thus untrustworthy, behaviours in operation, e.g. when the system encounters unseen scenarios or corner cases that were missing or under-represented in the training and verification datasets, or intentional manipulations exploiting brittle AI models. Fig. 1 shows a high-level overview of the AI training and deployment pipeline within the AD domain.

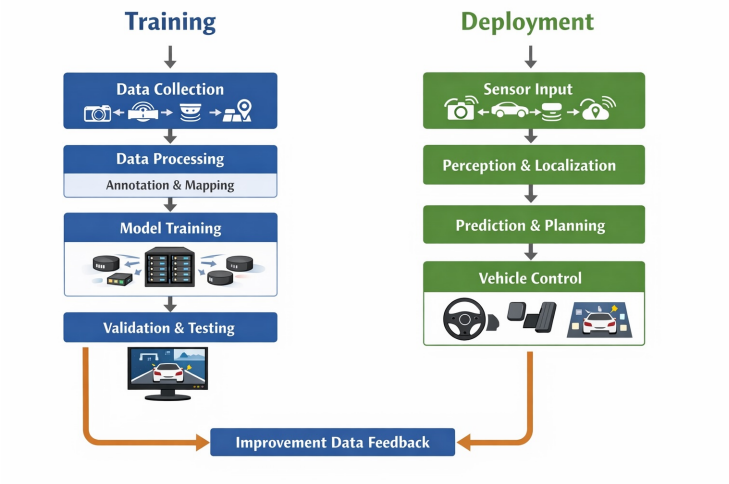


Fig. 1: A clear separation between training and deployment of AI models with respect to autonomous driving simplifies collection and training. However, it can open up mismatches between the training and deployment scenes hence derived datasets views from deployed/operations.

In safety-critical systems, in this case automated vehicles, risks result in human injuries, and may also trigger damaging, dramatic media coverage. Although many thousands die and are injured on roads worldwide, incidents involving automated vehicles make headline news, often globally. Therefore, ensuring safety requires methods and strategies to detect, mitigate and continuously manage potential failures across the entire lifecycle of an AI system.

We therefore need an AI safety assurance process and relevant approach to (i) understand the source of uncertainty

Co-funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Climate, Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them. Project grant no. 101069576.

by formally decomposing it into manageable types, (ii) define an uncertainty management strategy and practical approach for both development lifecycle phase and operation and monitoring stage.

Our key contribution in this paper is to bridge the gap between safety, security, physical integration of AI in critical systems with respect to *standards* and *common practices*, with a focus on data management activities and safety-critical autonomous driving systems. Specifically, guidelines, code and data have been made available through a project, ROADVIEW, which we elaborate on, and exemplify with, below.

In the autonomous driving domain, perception data typically consist of sensor readings from cameras, radars, lidars and corresponding annotations describing the surrounding traffic scenarios. Ensuring safe driving requires that safety-driven metrics be continuously evaluated in alignment with the system's defined safety scope as shown in Fig. 2.

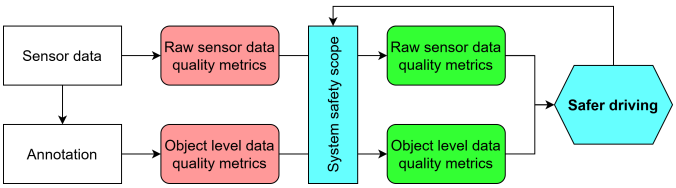


Fig. 2: Safety-driven data quality metrics in autonomous driving. Two paths through the process with feedback in the process.

II. Related Work

ISO/IEC 23894 *Artificial intelligence — Guidance on Risk Management* states that training, test and production data should be fit to the intended behaviour with respect to data type and quality [2]. In addition to this, the ISO/IEC 5259 standard series provides a unified framework that defines, measures and governs data quality across the data lifecycle to ensure reliable machine learning (ML) outcomes [3]. ISO/PAS 8800 further provides dataset related safety properties, such as accuracy, completeness and representativeness [4]. Safe AI development standards and practices e.g. EASA [5], AMLAS [6], AI-FSM [7] explicitly introduce data management as a separate phase within the development process. These frameworks emphasize that data is central to both AI development and operation, and there is a need for operational monitoring to detect when conditions may undermine system trustworthiness and safety. Standards such as ISO 21448 [8] and ISO/PAS 8800 provide partial data management guidelines [9]. These standards and practices together deliver a common message: Data management is fundamental to developing and deploying trustworthy AI in safety-critical applications. [10] discusses safety and AI, from feature selection to the engineering phase. This is not unlike this paper, however, we are a little more data-focused.

A paper that introduces a unified framework covering robustness, interpretability, controllability, and ethicality, and organizes alignment research into forward alignment, assurance, and governance is [11]. It serves as a key reference for system-level AI safety and alignment research. Amershi et al. present a set of empirically grounded design guidelines for effective human–AI interaction, synthesizing insights from HCI and AI system design [12]. The work defines actionable principles for transparency, control, feedback, and trust calibration, and has become a foundational reference for human-centered AI system design.

On the topic of autonomous driving and data quality, Liu et al. postulate that AD dataset surveys are not sufficiently extensive [13]. Therefore, the authors present an exhaustive study of 200+ datasets, with 350+ references, encompassing sensor modalities, varying data sizes, tasks, and contextual conditions. Of relevance is that they introduce a metric for evaluation akin to our DRLs, section V-C below. Our work differs from Liu in concrete scene-differences from training to operations using our comprehensive 5TB of project data.

From 2020, Vargas et al. look at autonomous vehicle sensors and their vulnerability to weather conditions [14]. From 2021, Yurtsever et al. present challenges, high-level system architectures, emerging methodologies, and core functions, including localisation, mapping, perception, planning, and human-machine interfaces [15]. The nuScenes publication, from 2020, contains a comprehensive survey of AD datasets [16] as well as their dataset. Bijelic et al. from 2020 examine adverse weather conditions and present a novel multimodal dataset acquired in more than 100km of driving in Northern Europe [17]. This dataset is the first large multimodal dataset encompassing adverse weather, with 100K labels for LiDAR, camera, RADAR, and gated near-infrared sensors. The Boreas dataset, from 2023, with 350 km driving data, was collected over a repeated route over one year, with seasonal variations and adverse weather conditions [18]. A list of AD datasets with harsh-adverse weather conditions is available [19].

III. AI challenges in safety-critical systems

The opaque and uncertain behaviour of AI models is the major challenge for their adoption in safety-critical applications. AI-based systems may produce unpredictable outputs when inputs deviate from their training datasets, often referred to as the model's generalization capability. However, in safety-critical contexts, uncertainty in generalization can lead to hazardous situations, making it difficult to guarantee safe performance across all operational scenarios within the system scope. Brando [20] proposed a formal decomposition of uncertainty into three main categories:

- **Domain:** AI models learn from a limited dataset that only approximates the true real-world distribution, creating a gap between what the model has been trained and tested on and the real scenarios it must handle in operation.
- **Epistemic:** AI models can only explore the limited solution space defined by their architecture and parameters. The gap between the optimized model within this space and the 'theoretically ideal' model results in epistemic uncertainty.

- **Aleatoric:** Captures the inherent variability in the model output, resulting from ambiguous or missing input information. Since the true ground-truth value is fundamentally non-existent in this case, the expected prediction should be represented as a distribution rather than a deterministic value.

This decomposition highlights that data quality, especially its representativeness, is a key uncertainty category. Managing the datasets used to train and verify an AI model is therefore as important as managing the trained model itself.

In safety-critical domains, safety-driven data quality becomes a primary factor in determining whether the system can operate safely within its intended scope. An example of such a safety-critical system is the vehicle perception system developed within the ROADVIEW project, Robust Automated Driving In Harsh Weather [21].

IV. Uncertainty management strategy

We propose a structured and systematic approach to manage and mitigate AI uncertainty by leveraging the above-mentioned decomposition and the guidelines in ISO 21448 - Road Vehicles - Safety of the Intended Functionality (SOTIF) [8]. SOTIF defines scenario categories based on two dimensions: known/unknown and hazardous/non-hazardous. During development, the goal is to maximize the *known* subspace, while during operation the goal is to ensure that the system operates within the known/non-hazardous subspace, minimizing the risk of accidents. The proposed data management strategy therefore focuses on domain uncertainty and consists of two main parts corresponding to the development lifecycle and the operation stage.

A. Domain uncertainty reduction in the development lifecycle

Domain uncertainty is identified and reduced during a dedicated data management phase within the lifecycle. The system's safety requirements are first allocated to corresponding data requirements. Based on these requirements, a structured set of data quality metrics will be derived, including both general Data Readiness metrics and safety-driven metrics specifically tailored to the system's identified hazards and mitigations. These metrics enable early identification of data quality gaps and guide additional rounds of data collection and preprocessing, such as augmentation and data cleansing. During this phase, the datasets and their expected safe boundaries are formally specified. System performance with respect to these boundaries will be documented together with the boundary specification. During the data management phase, the following topics are covered:

- **Data quality gap analysis**: This activity consists of (i) Completeness analysis to identify missing important features or data samples. Data balance analysis to baseline data distribution and identify gaps, (ii) Relevance analysis such as prototypes or criticisms to identify representative or unrepresentative data samples, and (iii) Accuracy analysis to examine annotation quality, consistency checks, distribution checks, and correlation analysis to assess how input features affect the labels.
- **Data quality gap closing**: Using data explainer techniques to guide data collection and preparation through: (i) uncertainty analysis to identify high-uncertainty regions where additional data would reduce the domain uncertainty, (ii) Diversity analysis, e.g. unsupervised data clustering, to verify whether data are adequately represented across diverse and representative regions.
- **Data augmentation**: Augmentation techniques using synthetic data or data reconstruction to close the identified distribution gaps. Feature importance analysis will guide the augmentation to close the gaps focusing on key features influencing safety-driven metrics. Identify realistic noises and adversarial patterns for augmentation targeting robustness requirements.
- **Data boundary specification**: Data descriptors trained during development are used to specify the datasets and also to validate input data later during the operation stage.

B. Control of residual domain uncertainty in operation stage

Residual uncertainty will be managed in operation through an operational safety architecture that encloses safety supervisors monitoring system boundary conditions to assess the trustworthiness of AI-based components outputs. If supervisors detect anomalous situations, such as anomalous input data, unusual model neuron activations or anomalous outputs, the decision-making component will be informed and mitigation actions will be triggered. Residual domain uncertainty is monitored during operation as follows:

- **Data anomaly detection**: Data descriptors trained and verified during development will be used as safety-driven data validators, together with other data quality validators, to assess whether incoming data are likely to belong to the *known* domain and thus the prediction of the AI models can be trusted for informed safe decisions.
- **Model anomalous behaviour detection**: Data anomaly detection models can also be trained to monitor models neural activations or model outputs to detect distribution shift or anomalous model behaviours not observed during development.
- **Rejected cases logging**: Operational situations where input data, interim extracted feature or output predictions are flagged as anomalous will be logged for diagnosis and future development cycles to improve the system or refine the specification.

C. Traceability - safety goals to data requirements/evidence

Traceability of the process starts with a clear definition of the system scope, including Operational Design Domain (ODD), the operational scenarios, and the intended functionalities. From this scope, the first activity to be conducted is Hazard Analysis and Risk Assessment (HARA), to identify hazards and derive system safety requirements. The system safety requirements will then be allocated to requirements for the data and the model, together with success criteria and

corresponding metrics. This creates the basis for assessing compliance, identifying gaps and generating evidence to support safety argumentation.

The allocation of system requirements to the data requirements is often guided by the key desiderata (such as those listed in [7]). For the ROADVIEW project, the following data requirements were specified:

- *Dataset completeness:* The dataset must include sufficient coverage of all object categories relevant to ODD and variations of weather conditions, sunny, rainy, snow, fog that may affect object recognition performance.
- *Dataset representativeness* of real-world scenarios, representative distribution of perception scenarios within each weather condition.
- *Data balance:* The dataset should maintain class balance under each weather parameter range.

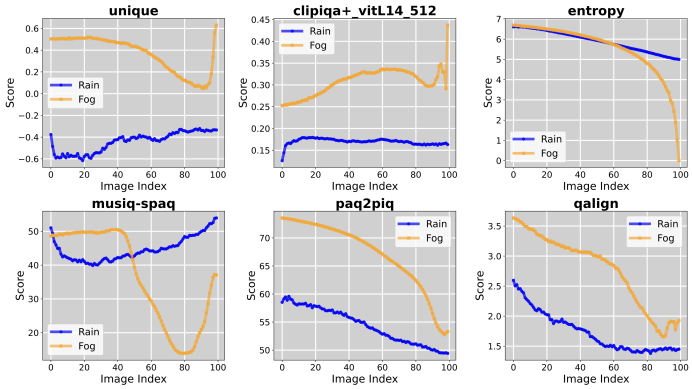


Fig. 3: Image metrics from six tools across a range of rain and fog impairments, see Fig. 5. 6 are shown here, 46 have been evaluated in [22], [23], Q-align performs best. Image quality metrics are an important part of the DRL, see the concept in Fig. 6. Data Readiness Levels are our method of assessing data quality, an important part of safety.

Sensor data are typically extremely high-dimensional, making it not practical to analyse every single dimension. To determine where the datasets meet the defined requirements, systematic methods (most often relying on Explainable AI techniques) are needed to identify and prioritize the key safety-driven metrics. These methods allow assessing dataset compliance, identify data gaps, guide gap closing actions and produce required evidence as part of the traceability requirements.

Since downstream AI models usually include one or more feature-extraction blocks followed by prediction modules that depend on those extracted features, it is of importance to assess whether the dataset actually provides the relevant features/insights needed for the intended functionalities and how these features are distributed across the dataset. To support this assessment and quantify the domain uncertainty, the following groups of data-explanation methods are selected:

1) **Data profiling:** evaluating data distributions and structure to confirm that the dataset meets predefined quality criteria. This step generates evidence (in forms of reports) on dataset characteristics, ensuring that no significant biases or inconsistencies affect model training and inference.
2) **Data prototyping:** extract representative examples from the dataset, ensuring that key patterns, objects, or concepts are well captured and understandable. This process enables explainability by linking AI decision-making to specific dataset elements.
3) **Data descriptors:** provide a structured way to summarize dataset characteristics, helping to define the boundaries between known and unknown data distributions. These descriptors support the identification of Out of Distribution (OOD) data. During the development cycle, it helps to specify and verify the dataset, while during the operation it can be used to formulate supervisory monitors to ensure that AI models operate within the known domain.

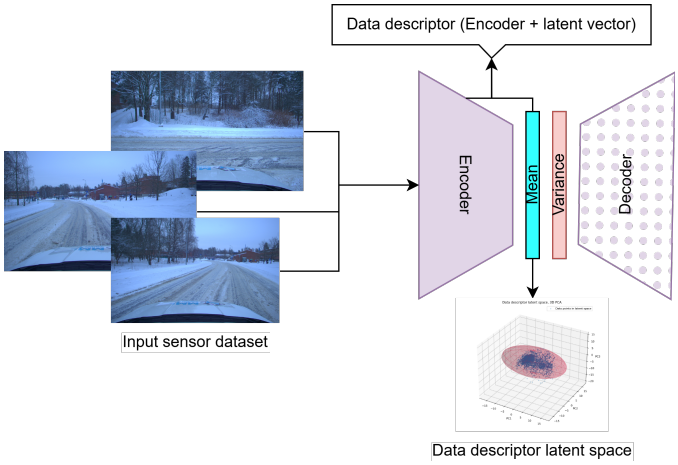


Fig. 4: Data descriptor example: A dataset is described by a train feature extractor model and corresponding feature space.

Fig. 4 shows an example of a data descriptor built using an AutoEncoder-based model. Each data point (i.e. image) is transformed by the encoder into a latent vector, a compact representation of most salient features of the data point. Collectively these latent vectors, corresponding to all samples in the dataset, form a subset of Euclidean space, reflect the structure of the dataset. The figure presents a three-dimensional visualization of the latent space, using PCA dimension reduction techniques. The dataset can then be described by the data descriptor, defined as the tuple Encoder, latent vectors, whereas the encoder specifies the projection transformation and the latent vectors represent the projected data.

V. Case Study: The ROADVIEW Project

An essential part of applying standards in practice is a Horizon Europe Innovation Action project, ROADVIEW [21]. To translate theory into practice, we implemented eight methods shown in Table I of how we worked with data to the

goal of safety-security-physical-AI in the realm of autonomous driving.

#	Measure	Purpose	Implementation
A	Data experience	Real & controlled	2 drives+1 track
B	Dataset augmentation	Add scene cases	Rain & fog
C	Data quality	Systemisation	DRLs
D	Data validation	'Debug' modalities	`rerun`
E	Data descriptors	Compact summaries	ML-based descriptor
F	Data profiling	Coverage	EDA
G	Data prototypes	Insights	Clustering
H	Safe operation	Control uncertainties	Safety supervisors

TABLE I: Data assessment using real sources.

A. *Datasets for AD training: 'Harsh weather included'*

We have access to 5TB of data gathered in a number of road scenarios, settings, day-night and weather conditions. Parts of the data collection are documented [24] whilst others are ongoing [21]. There is a strong focus on adverse, harsh, inclement weather conditions. In addition to system inclusion and testing, weather conditions impact this work by testing the ODD limits see Fig. 13.

Dataset (Environment)	Provider	Data Size	Details
FGI Open Road.	Finnish Geospatial Research Institute FGI. FI.	14 GB	1 urban & 1 rural journey, 49 km, from 03-12-2023.
REHEARSE Test Track.	Technological Institute of Ingolstadt, DE Cerema Inst. FR.	320 GB	2 testing grounds. H_2O-generated (Not snow/ice) 2023-ongoing
AVL Open Road.	AVL Software Functions, GmBH, DE.	2.5 TB	Open road drive in Southern Germany. 2023-2025

TABLE II: The three datasets utilised in this work.

As an example of the ideas in this paper we implemented them into the ROADVIEW project [21]. We used data from two open road drives and one test-track setup, see Table II. The ROADVIEW project has two open road drives and 1 test track plus a number of adverse weather scenarios. An up-to-date list of datasets that consider adverse weather conditions is available online, available from the link datasets.

B. *Data Augmentation: 'Filling the gaps'*

Data gathering and collection can only capture the conditions at *that time*. Whilst representative, the settings might not be sufficient to cover all cases. For example, a clear road is not the same as drizzle, fog, sleet, or snow. Despite having 5TB of data, controlled 'noise' was needed. Noise means rain and fog for images, and fog for point clouds. As well as the controllability in datasets, the REHEARSE test-track dataset has the possibility to generate real rain, drizzle, but not really fog, at least that remains in an outside environment. Over 3 million frames and 50 thousand point clouds without augmentation were rated, in addition to the samples impaired or augmented with weather conditions discussed here.

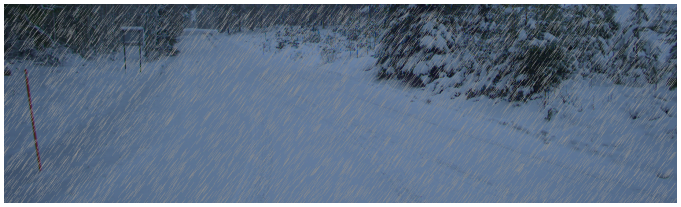


Fig. 5: Rain generation using the Albumentations package on the FGI dataset, see subsection II below. To calibrate the DRL Level 6, the impairments were controlled or varied, sometimes called 'noise modelling'. This particular image has been rain-impaired at a level of 20, on a scale from 0-100.

C. *Data Readiness Levels: 'Systemisation made easier'*

Data quality systemisation using a Readiness Level concept. Akin to Technical Readiness Levels (TRLs), Fig. 6 shows nine levels with respect to data quality within automated driving. An initial description of Readiness Levels was first introduced by Lawrence in [25]. He used 3 bands and classes within, whereas we kept to closer to the TRL levels in this scope. Noteworthy is the upward dependence of the levels, that a dirty or the effect of misaligned sensors will propagate up to the machine learning phase and ultimately object detection and therefore safety. The DRLs are organised to follow the *data flow*. Levels 6 and 7 are in the broad coverage of this paper.

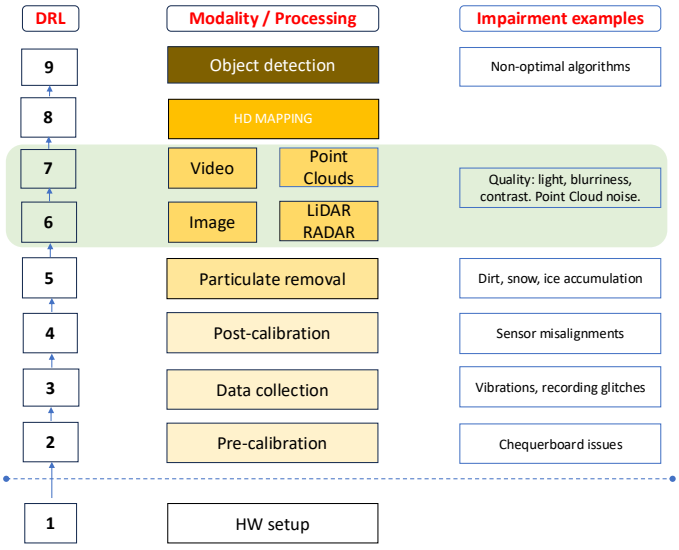


Fig. 6: *Leftmost column:* Data Readiness Levels. Used during training, data flows from DRL 2, calibration, upwards to object detection, DRL 9. *Central column:* indicate discrete steps aligned with the DRLs. *Rightmost column:* example impairments. The shaded green section indicates this paper's focus.

Each modality is handled separately and then combined into

a single scalar value, see [22][1].

D. Data validation across modalities: "Debugging the data"

Investigating scenes requires a tool to start, stop, fast forward and rewind specific portions. Identifying the *same* points in images and point cloud monitors how data varies over scenes. An example is in Fig. 7, for rerun see [26] and a video[2]. Inspection, or debugging, akin to software engineering, is a sanity check for data.

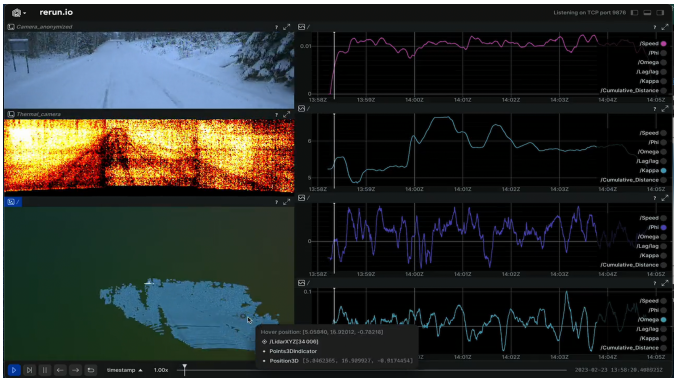


Fig. 7: Cross modality data validation in the rerun visualisation [26]. RGB, Thermal and point clouds are shown to the left and DRLs as continuous time series to the right.

E. Data profiling: 'Controlled coverage'

Data profiling involves analysing dataset distributions and structural properties, using Exploratory Data Analysis (EDA) techniques, to judge compliance with data requirements. This activity generates evidence in the form of reports on key dataset statistics aligned with the dimensions defining the system scope, i.e. ODD and operational scenarios. Data profiling is used to analyse both raw sensor data and annotation statistics. Fig. 8 shows the distribution of labelled object coordinates in ROADVIEW's FGI dataset. Based on this distribution, sparsity measures highlighting under-represented regions with too few data samples are illustrated in Fig. 9.

F. Data descriptors: 'Compact summaries'

The use of data descriptors is twofold: (i) To characterize the datasets during the development lifecycle and define the boundaries, and (ii) to provide a foundation for building anomaly detectors within the reference safety architecture.

Fig. 10 illustrates how an ML model can characterize a dataset and define its boundary. The example is based on the AVL dataset. We trained a Variational Auto Encoder [27](VAE) model to compress each image in the dataset into a compact numerical vector. Each image thus corresponds to a point in the 128-dimensional space that captures the most important visual features learned by the VAE model.

[1]Our contributions are available from GitHub https://github.com/roadview-project/No_reference_image_and_point_cloud_quality

[2]An AD sequence using DRLs is in this video sequence rerun.

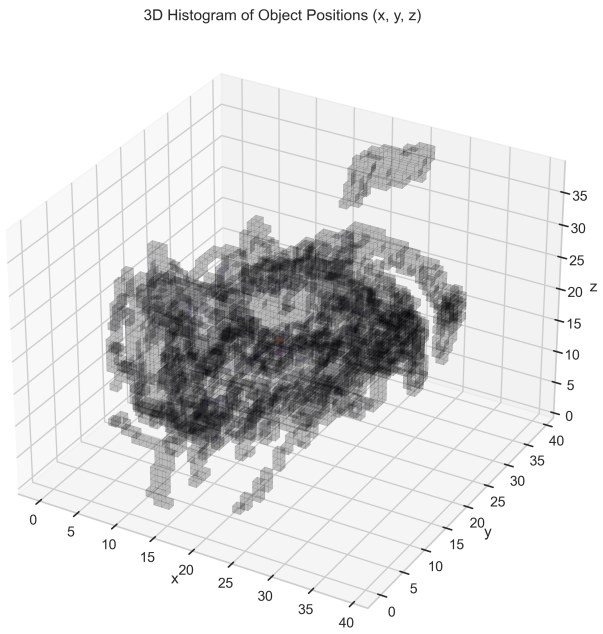
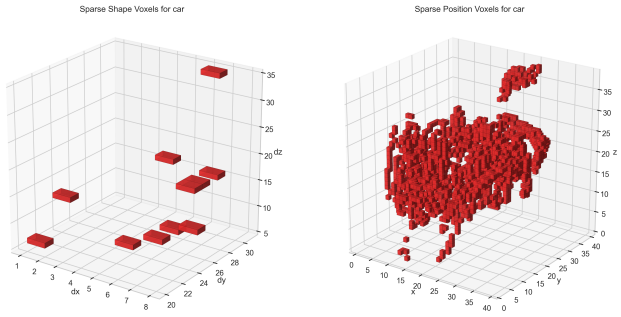


Fig. 8: A 3D histogram of annotated object positions in the FGI dataset. It highlights the distribution patterns and guides augmentation of important but under-represented interval in a discrete distribution, i.e. a histogram.



(a) Bounding box shape sparsity (b) Position sparsity

Fig. 9: The sparsity of positions and shapes for a car in an open-road journey from the Finnish (FGI) dataset detailed in Table II.

With respect to visualization, this latent space is projected into three-dimensions using Principal Component Analysis. The pink ellipsoid highlights the dataset boundary where most of the data samples are located. The introduction of data descriptors enables the formal specification of high-dimensional sensor datasets and defines their boundaries during development. The same model is later reused in the operation safety architecture, to assess whether incoming data falls within the safe limit, or not.

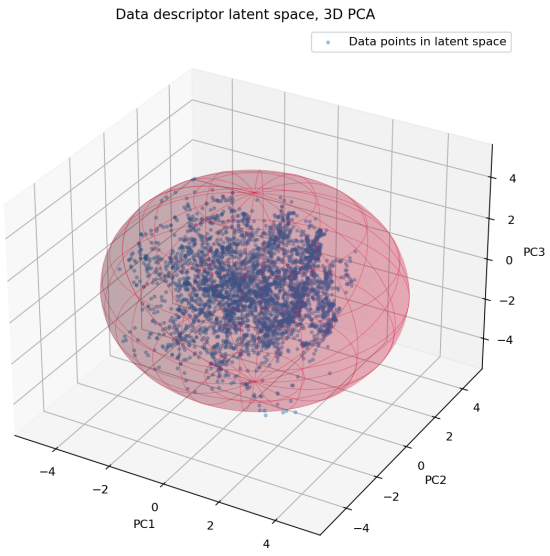
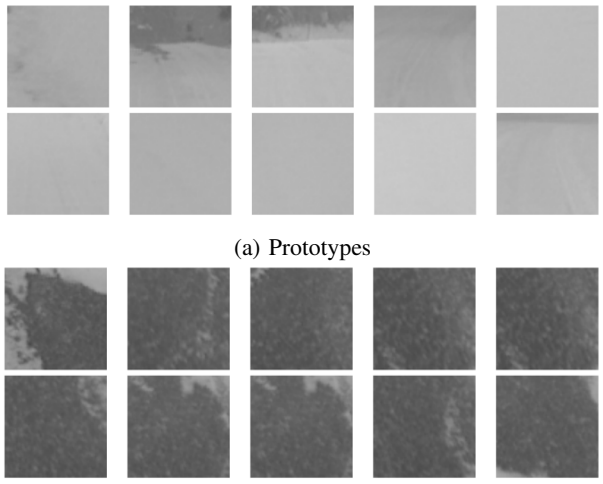


Fig. 10: FGI dataset distribution in the data descriptor's latent space and its estimated known boundaries shown within the ellipsoid.



(a) Prototypes

(b) Criticisms

Fig. 11: A set of in-distribution representative patches, the prototypes, and out-of-distribution representative patches, the criticisms, extracted from the FGI slipperiness dataset.

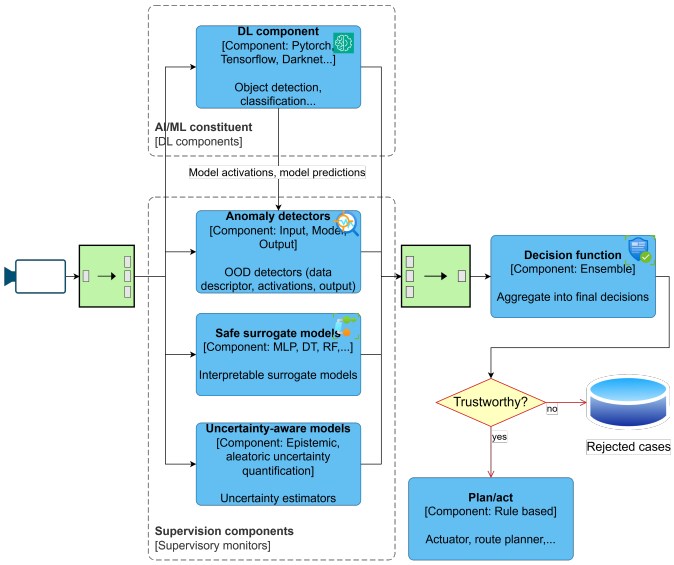


Fig. 12: A reference safety operational architecture. Shown is a set of supervisory components to monitor systems operations. Both boundaries and uncertainties are monitored for safe-limit violations.

G. Data prototypes: 'Domain insights'

Prototypes and criticisms are data-summarization methods that identify representative examples, called *prototypes*, and non-representative examples, called *criticisms*, that characterize a dataset. These methods evaluate whether the dataset captures the key features that the AI model will later rely on. The use of prototypes and criticisms thus helps to quantify and reduce domain uncertainty by revealing how well the dataset represents the operational domain and aligning data quality assessment and gap analysis with domain-specific insights.

An example of computed prototype patches (a) and criticism patches (b) for ROADVIEW's FGI slipperiness dataset [28] is shown in Fig. 11. We first extract image features in small patches and compute prototypes, which are representative samples and criticisms, for samples that deviate significantly from the norm. For each pixel with a ground-truth label, a 70×70 pixel patch centred on the pixel was extracted, empirically. Unsupervised clustering was applied in order to group similar patches. The centroid of each cluster is selected as a prototype, while patches whose distance to their nearest centroid exceeds a threshold that are selected as criticisms.

H. Safe operation: 'Control uncertainties'

To continually monitor residual domain uncertainty and prevent it from leading to unreasonable hazardous situations, we employ the safe operational architecture illustrated in Fig. 12 where the corresponding supervisors reuse the data descriptors introduced in Section V-F.

The ML-based data descriptor trained on the FGI dataset during development is reused by a supervisor component in operation. The chosen ML model is a Variational Auto Encoder (VAE), whose encoder uses convolutional layers followed by fully connected layers to produce the mean and log-variance of a compact latent vector. The decoder mirrors this architecture to reconstruct the input. During operation, this VAE model acts as a supervisor that validates whether the incoming input data stay within the boundaries captured during development. Fig. 13 illustrates this process. Each row shows the input image, its reconstructed version, and the anomaly score, which is computed as the L2 reconstruction error. The

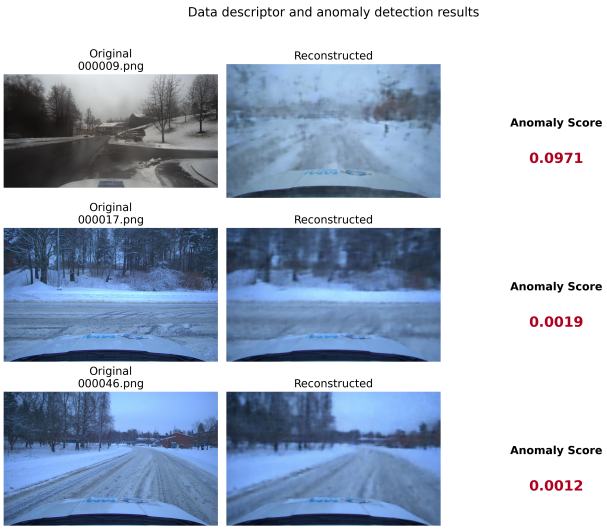


Fig. 13: Data descriptors trained on the FGI dataset. Anomaly scores are computed as the distance between the input image and its reconstruction. In the first row, the input image is from another dataset. Numerically, the anomaly score is 51% and 81% higher compared to the other two scenes, thereby exceeding the predefined admissible threshold, and hence indicates an outlier.

admissible threshold is chosen so that the captured anomaly scores computed for the verification dataset are all considered acceptable and within the expected variation of in-distribution samples. In the first row, the input image is taken from another ROADVIEW dataset, the AVL dataset as shown in Table II. This example image has a much higher anomaly score, exceeding this threshold, indicating an outlier outside the *known* region and therefore corresponds to increased domain uncertainty.

VI. Discussion

Drawing on comparisons with the software engineering industry, we are trying to improve both the data and algorithms. In traditional algorithm development, the data is considered to be well-known, or manageable, e.g. not dividing by zero, or using the median where data is skewed. In this case, the data is not so clearly defined, delineated or completely described. In the AI age, the data and algorithms work in conjunction, and therefore the data is equally as important as the algorithms. This represents a significant shift in data engineering, computer science, statistics.

As the data changes sector to sector, or case to case, one needs specific tests, albeit using general principles, with best practices, and this has been the thrust of this work. Principles, standards, definitions and well-defined domains are welcome, but we need best practices and experience as well. We have applied practices using automated driving and *its data* as a case study, with an aim to add common sense, safety and sanity checks as well as some safeguards in the field of autonomous driving.

Security is not really addressed in this work, as we do not consider any kind of adversary. Naturally, this introduces a new set of safeguards, including complete testing of those in place and strict adherence of those implemented. The most important aspect in a security setting is to raise some kind of alert, warning or even 'stop' should the safeguards be violated in operation, and constructing correct ones in training.

We would like to point out this is very much ongoing work, as is the domain of high quality data(sets) for automated driving. Perfection is not possible, but a series of safeguards is needed from paper to practice, from data to models, therefore we will continue in this vein to ensure safety for physical AI systems.

VII. Future work

This paper shows the relevant standards and issues within autonomous driving with respect to safety and physical AI. Security is not discussed, but is earmarked as future work. We are still closing the gap between 'desk research' and code + real data. We are constructing feedback to the standards organisations in terms of inconsistencies, ambiguous or vague formulations or improvements to help produce or hone well-defined and implementable standards, both formally and with industry partners.

Data augmentation was only employed for weather conditions, rain and fog, but more should be covered, drizzle, sleet, snow etc[3]. Augmentation should be extended to cover other road cases, extra vehicles, glare from oncoming headlights, potholes, roadside detritus, and other potential hazards.

Computer-generated augmentations offer flexibility and cost savings compared with practical data gathering or test track setups, but should be available to validate the augmentations. In terms of quality, image impairments are relatively well known, point clouds less so. In a reference situation where one can compare real with estimated scenarios, i.e., distances, and known objects, one can 'rate' point cloud quality. An assessment of the two most cited no reference point cloud algorithms can be found in [22].

Processing large quantities of data, in raw and augmented modes requires considerable processing. We are looking at these costs, both time and financial with tools such as weights and biases `wandb.com`.

VIII. Conclusions

Few sectors demand stricter safety guidelines, standards and systems other than the automotive one. With an eye on safety, this paper has looked at sensors, impairments, data quality and its impact on downstream machine learning. Poor data requires additional cycles of the drive-collate-process-assess process. A complicating factor is adverse weather, which in this work we factor in with respect to Nordic conditions. We introduced data readiness levels, akin to TRLs for data, data descriptors and implementation of explainable AI. We also use a tool for checking the modalities *at the same time* for

[3]Indeed, these are *Nordic* harsh conditions, but equatorial and tropical zone harshness conditions exist namely sand, heat haze etc.

AD, separate from the processes above, to *debug the data*. Our contributions were a systematic survey of data management for safety-critical systems, namely automated driving. We focused on autonomous *driving*, that is the scenes and weather for vehicles, rather than the vehicles solely.

IX. Acknowledgements

The authors would like to thank Heikki Hyyti at the Finnish Geospatial Institute (FGI), Yuri Poledna at the Carissma Institute at the Technical Institute of Ingolstadt, Germany and Stefan Parzefall at AVL Germany in providing data for our use case. We are also indebted to Fredrik Warg at The Research Institutes of Sweden, RISE AB, for his invaluable comments on this manuscript.

References

[1] S. Shafaei, S. Kugele, M. H. Osman, and A. Knoll, "Uncertainty in machine learning: A safety perspective on autonomous driving," in *Computer Safety, Reliability, and Security: SAFECOMP 2018 Workshops, ASSURE, DECSoS, SASSUR, STRIVE, and WAISE, Västerås, Sweden, September 18, 2018, Proceedings*. Berlin, Heidelberg: Springer-Verlag, 2018, p. 458–464. [Online]. Available: https://doi.org/10.1007/978-3-319-99229-7_39

[2] International Organization for Standardization, "Information technology — Artificial intelligence — Guidance on risk management (ISO/IEC Standard No. 23894:2023)," 2023. [Online]. Available: https://www.iso.org/standard/81230.html

[3] ——, "Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 1: Overview, terminology, and examples (ISO/IEC Standard No. 5259-1:2025)," 2025. [Online]. Available: https://www.iso.org/standard/81088.html

[4] ——, "Road vehicles — Safety and artificial intelligence (ISO Standard No. 8800:2024)," 2024. [Online]. Available: https://www.iso.org/standard/83303.html

[5] M. Consortium, "EASA Research – Machine Learning Application Approval (MLEAP) Final Report," EASA, Tech. Rep., May 2024. [Online]. Available: https://www.easa.europa.eu/sites/default/files/dfu/mleap-d4-public-report-issue01.pdf

[6] R. Hawkins, C. Paterson, C. Picardi, Y. Jia, R. Calinescu, and I. Habli, "Guidance on the Assurance of Machine Learning in Autonomous Systems (AMLAS)," *arXiv:2102.01564 [cs]*, Feb. 2021, arXiv: 2102.01564. [Online]. Available: http://arxiv.org/abs/2102.01564

[7] J. Fernández, I. Agirre, L. Belategi, J. Pérez-Cerrolaza, A. Adell, J. Imaz, C. Donzella, and J. Abella, "AI-FSM," Apr. 2024, version Number: v1.0. [Online]. Available: https://doi.org/10.5281/zenodo.10964402

[8] International Organization for Standardization, "Road vehicles — Safety of the intended functionality (ISO Standard No. 21448:2022)," 2022. [Online]. Available: https://www.iso.org/standard/77490.html

[9] V. J. Expósito Jiménez, B. Winkler, J. M. Castella Triginer, H. Scharke, H. Schneider, E. Brenner, and G. Macher, "Safety of the intended functionality concept integration into a validation tool suite," *Ada Lett.*, vol. 43, no. 2, p. 69–72, Jun. 2024. [Online]. Available: https://doi.org/10.1145/3672359.3672369

[10] X. Li, P. Ye, J. Li, Z. Liu, L. Cao, and F. Wang, "From features engineering to scenarios engineering for trustworthy ai: I&i, c&c, and v&v," *IEEE Intelligent Systems*, vol. 37, no. 4, pp. 18–26, 2022.

[11] J. Wang *et al.*, "Ai alignment: A comprehensive survey," *arXiv preprint arXiv:2310.19852*, 2023. [Online]. Available: https://arxiv.org/abs/2310.19852

[12] S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. N. Bennett, K. Inkpen, J. Teevan, T. Kulesza, and E. Horvitz, "Guidelines for human-ai interaction," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 2019, pp. 1–13.

[13] M. Liu, E. Yurtsever, X. Zhou, J. Fossaert, Y. Cui, B. L. Zagar, and A. C. Knoll, "A survey on autonomous driving datasets: Data statistic, annotation, and outlook," 2024.

[14] J. Vargas, S. Alsweiss, O. Toker, R. Razdan, and J. Santos, "An overview of autonomous vehicles sensors and their vulnerability to weather conditions," *Sensors*, vol. 21, no. 16, 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/16/5397

[15] E. Yurtsever, J. Lambert, A. Carballo, and K. Takeda, "A survey of autonomous driving: Common practices and emerging technologies," *IEEE Access*, vol. 8, pp. 58 443–58 469, 2020. [Online]. Available: http://dx.doi.org/10.1109/ACCESS.2020.2983149

[16] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020, pp. 11 618–11 628.

[17] M. Bijelic, T. Gruber, F. Mannan, F. Kraus, W. Ritter, K. Dietmayer, and F. Heide, "Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 11 679–11 689, 2020.

[18] K. Burnett, D. J. Yoon, Y. Wu, A. Z. Li, H. Zhang, S. Lu, J. Qian, W.-K. Tseng, A. Lambert, K. Y. Leung, A. P. Schoellig, and T. D. Barfoot, "Boreas: A multi-season autonomous driving dataset," *The International Journal of Robotics Research*, vol. 42, no. 1-2, pp. 33–42, 2023.

[19] I. Marsh, "Autonomous driving datasets and weather," Nov. 2025. [Online]. Available: https://ianmarsh.org/autonomous-driving-datasets

[20] A. Brando, I. Serra, E. Mezzetti, F. J. Cazorla, and J. Abella, "Standardizing the Probabilistic Sources of Uncertainty for the sake of Safety Deep Learning," in *Proceedings of the Workshop on Artificial Intelligence Safety 2023 (SafeAI 2023) co-located with the Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI 2023), Washington DC, USA, February 13-14, 2023*, ser. CEUR Workshop Proceedings, G. Pedroza, X. Huang, X. C. Chen, A. Theodorou, J. Hernández-Orallo, M. Castillo-Effen, R. Mallah, and J. A. McDermid, Eds., vol. 3381. CEUR-WS.org, 2023. [Online]. Available: https://ceur-ws.org/Vol-3381/11.pdf

[21] ROADVIEW Project, "ROADVIEW – Robust Automated Driving in Extreme Weather," https://roadview-project.eu/, accessed: 2025-07-11.

[22] M. Ian, S. Victor, P. Yuri, H. Heikki, S. Martin, and E. Eren, "Data readiness levels for automated driving," in *To appear, Intelligent Vehicles, April 2026, Detroit, USA*, 2026.

[23] V. Stenmark, "Evaluating NR-IQA & NR-PCQA Methods on Weather-Distorted Data in Autonomous Driving," Master's thesis, KTH, School of Electrical Engineering and Computer Science (EECS), 2025.

[24] Yuri Poledna et. al, "adveRse wEatHEr datAset for sensoRy noiSe modEls," 2024. [Online]. Available: https://s3.ice.ri.se/roadview-WP3-Warwick/T3.2%20-%20Create%20Dataset/rehearse/index.html

[25] N. D. Lawrence, "Data readiness levels," 2017.

[26] Rerun Development Team, "Rerun: A visualization sdk for multimodal data," Online, 2024, available from https://www.rerun.io/ and https://github.com/rerun-io/rerun. [Online]. Available: https://www.rerun.io

[27] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," *arXiv:1312.6114 [cs, stat]*, May 2014, arXiv: 1312.6114. [Online]. Available: http://arxiv.org/abs/1312.6114

[28] J. Maanpää, J. Pesonen, I. Melekhov, H. Hyyti, and J. Hyyppä, "Road Grip Uncertainty Estimation Through Surface State Segmentation," in *Image Analysis*, J. Petersen and V. A. Dahl, Eds. Cham: Springer Nature Switzerland, 2025, pp. 231–244.