NVIDIA Halos Safety Evaluation Framework

By: Sever Topan US Anitha Anand US Joerg Rossow DE Jonathan Mcneely US Viola Wu US Apoorva Sharma US Martijn Tideman Matteo Oldoni IT Jonas Nilsson US Riccardo Mariani IT

Introduction

Demonstrating AV safety requires proving performance at rates on the order of one failure every 10^7 to 10^8 hours of driving – a volume no pre-production fleet can drive and no single tool can address. This demands a data-driven combination of on-road testing, simulation-based evaluation, and production fleet monitoring. Automotive standards are increasingly formalising exact safety requirements as the industry shifts to AI-based architectures.

This document describes the NVIDIA Halos Safety Evaluation Framework (SEF) – an ecosystem of tools and accompanying guidelines for generating sufficient evidence to support AV safety cases across autonomy levels, from L2 active safety systems to L4 robotaxi autonomy. NVIDIA Halos SEF builds on more than 330 research papers and 1,000 patents developed within NVIDIA Halos OS.

Our mission is twofold:

- We aim to draw out a specific and actionable recipe of the necessary components for standards-compliant safety evaluation, from L2 to L4 autonomy.
- We aim to offer a select number of these components to the market, such as NVIDIA NuRec and Cosmos Transfer. As many ecosystem partners also specialize in individual components, we encourage partnership and collaboration to implement Halos SEF as a complete solution.

Safety evaluation is not achieved by a single silver-bullet technology, but is a combination of data, software, and processes that come together to create statistical evidence. In this document we will share a high-level overview of NVIDIA Halos SEF.

Our Architecture

The engineering challenge behind AV Safety Evaluation stems from the stringency of the safety targets that need to be proven. Most AV safety performance targets are grounded in human benchmarks, and require evidence that the system will not result in a fatality more often than once every 10^7 to 10^8 hours of driving[1]. Naively testing this volume of hours in-car is infeasible. Simulation technology is thus key to establishing these performance bounds, though it immediately raises the question of simulation fidelity.

Volume aside, we face operational challenges of data collection, labeling and curation at vast scale. We must provision data centers to store petabyte-scale driving logs and execute our simulators on them. We must sift through thousands of performance metrics to identify and prioritize safety issues, ground-truth errors, and simulation realism gaps. We contend with questions of how to effectively guard against overfitting in our pipeline, and how to account for compounding errors from many chained automated tools.

Halos SEF provides specific and actionable guidelines, along with industry-leading tools that resolve these issues. It allows for the measurement of Key Performance Indicators (KPIs) that provide evidence to support the automotive safety case. It comprises three primary components:

- **Process Governance**: A safety lifecycle management process for requirements and data-driven risk assessment of KPI results
- The **Data Foundation**: a set of tools for creating, labeling, curating, and augmenting datasets
- The **Simulation Core**: a variety of simulators, automated analysis software, and a result visualization platform for generating safety evidence.

The overall safety framework is designed to be cyclic and iterative, where past findings and releases inform new requirements for data, labeling simulation, and analysis. Halos SEF is a flywheel that generates evidence to support automotive safety cases. We'll now trace a path through Halos SEF, walking through each component in detail.

[1] Human fatality rates occur at a rate between once every 10^6 to 10^7 hours of driving, depending significantly on the geographic region in question. The specific target that autonomous vehicles must be held at is subject to debate, but a rough factor of 10 is often applied over top of the human fatality rate to account for impaired driving and other uncertainties. For the purposes of this document, the targets aim to illustrate the magnitude of the evaluation infrastructure necessary to demonstrate the safe deployment of autonomous vehicles.

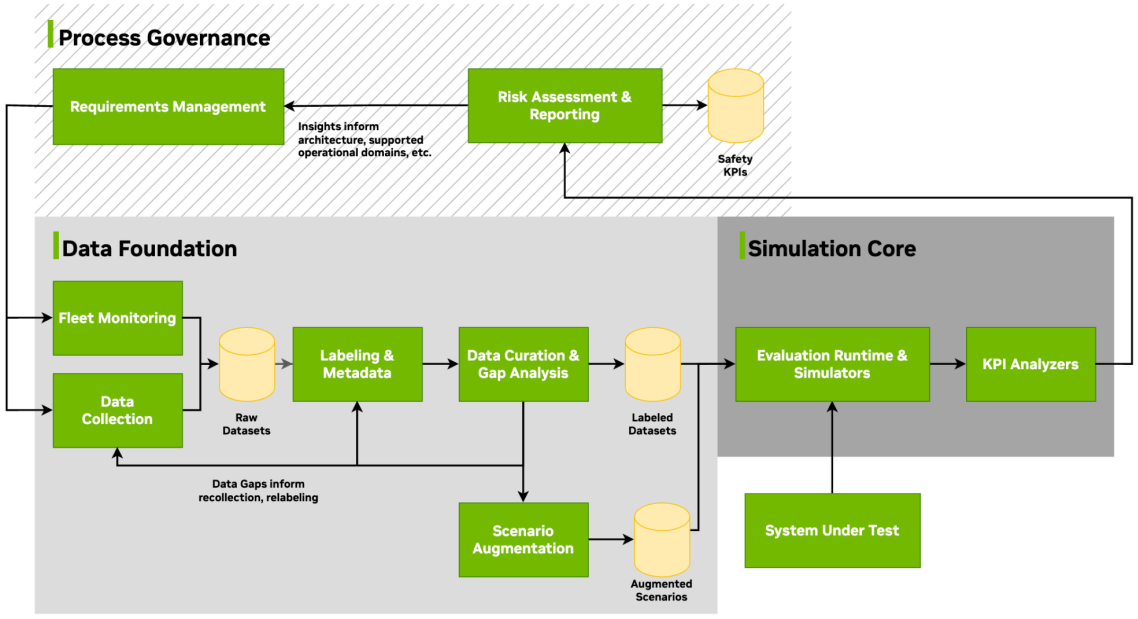

Figure 1: The major components of Halos Safety Evaluation Framework.

Process Governance: Requirements Management

The core goal of Halos SEF is to provide statistical guarantees confirming that the system adheres to its safety goals. Simply put, it works on the principle that a safety requirement must not only specify a necessary system property but also define the objective level of certainty required for its validation. As such, the definition and management of these safety concepts and requirements are the starting point of Halos SEF.

Figure 2: Process Governance within Halos SEF, focusing on safety requirements management processes.

Safety requirements are derived through a first-principles, holistic safety-case-driven approach. We anchor our analysis in the application's ODD to identify both system malfunctions and risks from environmental triggers, known performance limitations and operational assumptions. These analyses are then translated into safety concepts that allocate safety requirements and acceptance criteria across the system, software, hardware, AI/data lifecycle, and operational controls. The allocated requirements define the technical measures, interfaces, monitoring, redundancy, fault-tolerant behavior, and V&V evidence needed to keep risk within acceptable bounds. For more details on deriving quantifiable safety KPIs and managing triggering conditions within AI and machine learning lifecycles, see Effective and reflective assurance for AI-based autonomy.

This is all brought together in a structured safety assurance argument, also known as the safety case, which demonstrates that residual risk has been brought down to an acceptable level. It is a living document that provides safety assurance not just at initial launch, but throughout the product life cycle. The entire safety case framework is built upon the foundational requirements of industry standards. Standards such as ISO 26262, ISO 21448/SOTIF, ISO PAS 8800, and UL 4600 provide the frameworks to structure and prove these assurance arguments in our safety case. TÜV Rheinland has successfully completed an independent safety assessment for the NVIDIA NDAS SAE Level 4 autonomous system's safety management planning. Built on the NVIDIA DRIVE AGX Hyperion 10 platform, the system's Functional Safety, SOTIF, and AI Safety Management planning — along with its V&V strategy — were assessed against ISO 26262, ISO 21448 (SOTIF), and ISO/PAS 8800 standards. Throughout the assessment, NVIDIA demonstrated relevant competencies on the required safety activities and safety-driven processes required for the next generation of autonomous mobility.

Figure 3 below illustrates a representative safety case argument structure in this context.

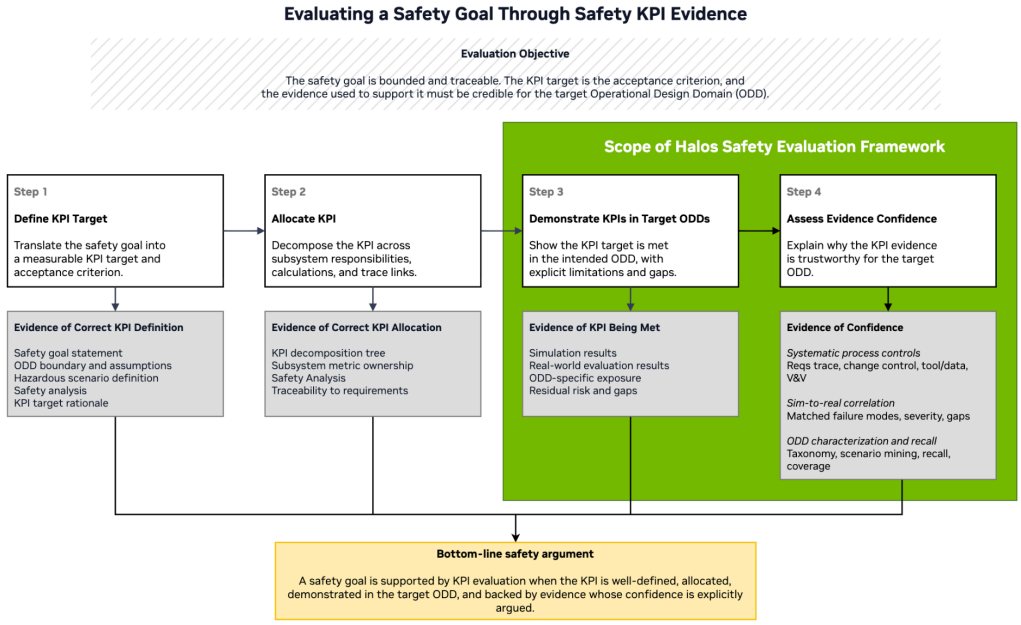

Figure 3: Representative safety case argument structure.

Each safety requirement can be thought of as tracing a particular path through our safety evaluation framework, touching only the elements necessary for demonstration. Different partners pursuing different autonomy levels may opt to invest in the paths most relevant to their business strategies. As we walk through the components of Halos SEF, we will use the safety requirements below as illustrative examples of how the different components come together to form a complete safety story.

Example Safety Requirement	**Autonomy Level**	**Example Safety Requirement Target**
The vehicle's AEB system shall trigger in the CCFTap NCAP scenario	L2	Pass all CCFTap NCAP scenario variations with additional fuzzing
The vehicle shall avoid AEB Ghost Brakes a high speeds	L2	1 failure every 10,000 hours
The vehicle shall avoid collisions during unprotected intersection traversals that may result in serious injury	L2++	1 non generally controllable traffic conflict every 20,000 hours of L2++ operation
The vehicle shall be free from camera obstacle perception errors relevant to lane changes	L4	1 failure every 10,000 hours

Table 1: A sampling of illustrative safety requirement examples.

Data Foundation

The Data Foundation of Halos SEF concerns itself with constructing the datasets necessary for safety evaluation, from data collection to labeling, curation, and augmentation.

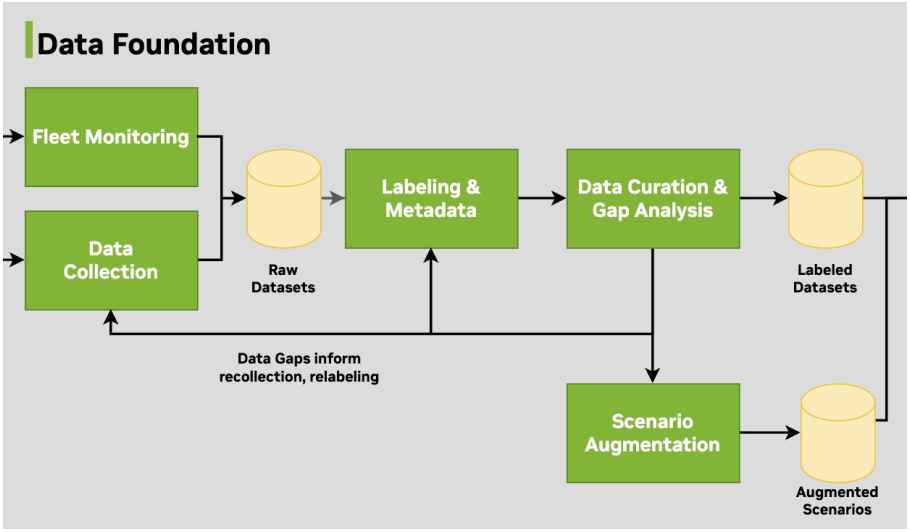

Figure 4: The Halos Safety Evaluation Framework Data Foundation.

Data Collection

The primary factors guiding data collection are the volume and distribution of the dataset in question. Demonstrating a 20,000-hour KPI target with 95% confidence requires roughly three times that volume of error-free driving. Furthermore, in order for a statistical argument based on this data to hold water, it must be sampled according to the expected customer distribution. A dataset that overrepresents highway miles, for example, cannot be used to make claims about urban intersection performance. Our fleet therefore requires dedicated tooling to both plan collection operations and to analyze the resulting data they produce, ensuring that volume and distribution targets are met before the data is released into downstream pipelines.

Labeling & Metadata

In order to perform data distribution analysis, as well as to enable many downstream SEF components, ingested data must be labeled appropriately. Different KPIs come with different labeling requirements, from obstacle cuboids to lane polylines, to weather metadata. NVIDIA has built a host of automated and human-supervised data labeling tools on top of the Hyperion sensor platform that allow us to create relevant labels at scale.

It is important to note that for any such tool, the precision and recall of the labels themselves must be characterized and accounted for in downstream safety arguments. A systematic issue in lane graph autolabels, for instance, can percolate throughout our safety claim and bias the final result.

As an example of the collection and labeling tooling that NVIDIA has built, earlier this year we open-sourced 1,700 hours of labeled data in the Physical AI Autonomous Vehicles Dataset.

Curation & Gap Analysis

60,000 hours of labeled driving logs are a significant amount of data to store and simulate. We note, however, that the L2++ safety requirement in our running example focuses on intersection performance. We may thus curate a significantly more efficient subset of that 60,000 hours for testing, focused on scenarios relevant to the safety requirement at hand. Tailoring datasets to the safety requirements often opens up opportunities for optimizing dataset structure and reducing simulation cost by orders of magnitude.

It is at this stage that we may also identify systematic gaps in our dataset that warrant recollection. NVIDIA leverages a host of traditional statistical methods for this task in conjunction with cutting-edge AI-based tools. Vector embedding search technologies such as NVIDIA Cosmos Dataset Search open up unique opportunities for semantic curation, allowing engineers to find not just driving logs that match on metadata, but logs that match on the underlying driving situation. Gaps identified at this stage feed back into the data collection plan, closing a loop between what the safety requirement demands and what the fleet goes out to capture.

Data Augmentation

While generating a dataset that represents 60,000 hours of intersection traversals is sufficient to prove our statistical performance bounds, certain scenarios may be reasonably foreseeable yet very unlikely to be sampled, falling in the infamous long tail of autonomous driving.

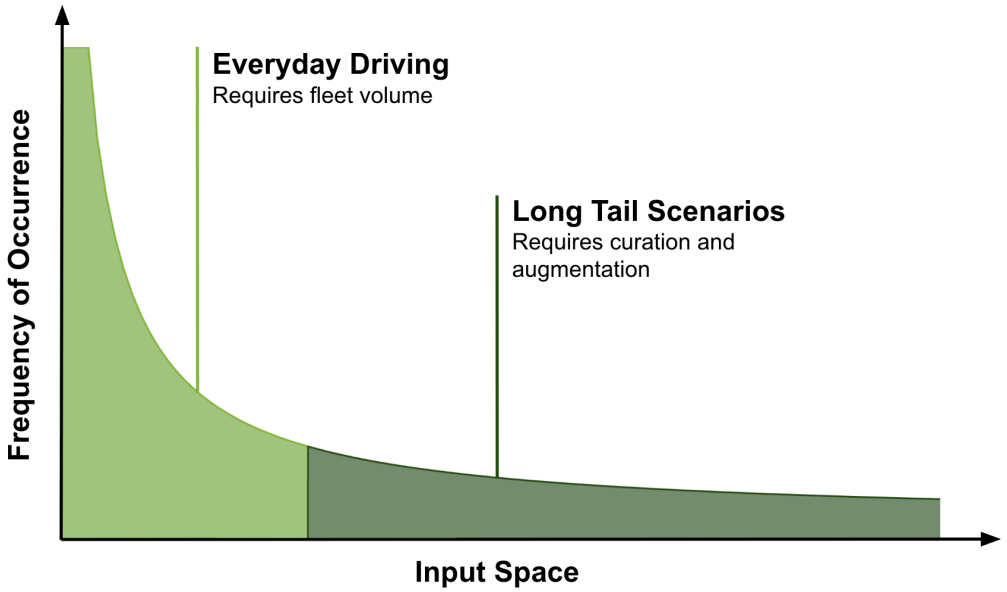

Figure 5: Halos SEF must account for both everyday driving, as well as the long tail scenarios; Brute-force data collection can struggle to uncover these rare yet dangerous scenarios

Long tail scenario specifications can be mined from literature and national crash databases, but are inherently rare and sometimes dangerous to collect. Data augmentation technologies are thus an excellent solution for long-tail scenario coverage.

Augmentation can take two forms: synthetic scenario generation and scenario perturbation. While both approaches are valid, at NVIDIA we focus on the latter. During scenario perturbation, an existing log is modified to produce variants of the original encounter – adjusting ego approach speed, shifting the timing of a cross-traffic actor, or introducing a completely new occluding vehicle at a critical moment. NVIDIA NuRec in particular allows us to leverage neural reconstruction to render the perturbed scene at high fidelity. Technologies such as Cosmos Transfer furthermore allow us to expand the operational domain of our testing by modelling environmental effects such as rain, snow, or wildfire smoke.

Figure 6: Cosmos Transfer for data coverage of diverse weather & lighting.

Real driving logs anchor our statistical claims to the true customer distribution, while augmented data ensures rare-but-foreseeable scenarios are represented with sufficient density. As with autolabels, the realism characteristics of our augmentation tooling must themselves be accounted for in downstream safety arguments. Once produced, augmented data enters our pipeline alongside real logs, forming a dataset ready for the Simulation Core.

Fleet Monitoring

Safety requirements need to be proven before a new feature is brought to market. For initial bring-up, this evaluation is typically performed using a dedicated and carefully monitored development fleet. Once a feature is released to the market, customer logs become an immensely valuable data source[2]. The customer fleet dataset enables post-release KPIs that answer questions such as: were our development-time assumptions about the customer operational distribution accurate? Are measured failure rates consistent with what the deployed fleet experiences? Can we find any scenarios that warrant targeted recollection or augmentation?

Customer datasets can cover orders of magnitude more driving mileage than a development fleet. At this scale, we must rely on the same Data Foundation tooling used elsewhere in Halos SEF to ingest, store, label, and curate incoming logs – sifting data streams down to the subsets relevant for each safety claim. Fleet monitoring turns the safety case from a one-time bring-up exercise into a continuously maintained property of the deployed system.

[2] Of course, privacy is a key concern whenever collecting and storing customer data, and any such pipeline must be architected around consent, anonymization, and region-specific regulatory requirements from the outset.

Simulation Core

Now that we have constructed a dataset to evaluate our system, we are positioned to simulate our system and produce the safety performance evidence that we are after.

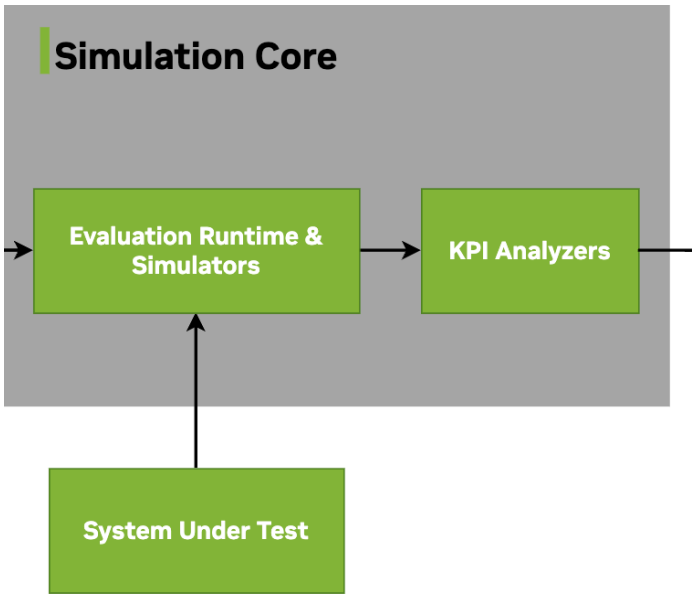

Figure 7: The Halos Safety Evaluation Framework Simulation Core.

Evaluation Runtime & Simulators

Different safety claims impose different demands on scale, fidelity, and cost of the necessary simulation. Halos SEF does not prescribe a single simulator. Instead, it provides a menu of simulation technologies, along with guidelines for choosing the right tool for the claim at hand. For our running L2++ intersection example, evaluating collision avoidance during unprotected turns, requires closed-loop simulation with high sensor realism: the system's actuation must influence the trajectories of surrounding agents, and the rendered sensor stream must be faithful enough that the perception stack behaves as it would in the real world. Contrast this with an AEB false-positive claim, where open-loop replay of recorded sensor data is not only sufficient but orders of magnitude cheaper – the mere presence of a ghost-brake trigger counts as evidence, and there is no need to model downstream vehicle dynamics. The space of possible simulation technologies becomes even more complicated as we start considering HiL & PiL setups where modelling correct behavior requires actual hardware components present in the production vehicle.

At NVIDIA, our most advanced simulators use NuRec – a neural reconstruction and rendering technology that produces photoreal, sensor-accurate recreations of real driving scenes. NuRec allows us to take a recorded log, reconstruct the 3D scene, and re-simulate it in a closed-loop with

neurally rendered novel view synthesis. As mentioned in our Data Augmentation section, NuRec allows us to perturb ego behavior and inject new actors in the scene, all while preserving the sensor realism that AI-based perception stacks demand. This capability is what makes claims like our L2++ intersection target tractable: we can take real intersection encounters, simulate novel ego trajectories through the scene, and measure collision outcomes at a scale that is challenging with pure synthetic simulation.

Regardless of which simulator is chosen, its realism must itself be characterized before its outputs can be trusted as safety evidence. A common technique in Halos SEF is a log-to-sim consistency check: we take a real-world driving log, hold every variable constant – the same vehicle build, the same software version, the same sensor inputs – and re-run it through simulation. Any divergence between what happened on the road and what the simulator reports is a measure of the simulator's realism gap. Sim realism should be continuously measured throughout the project lifecycle, as both the simulator and vehicle software evolve. These gaps are tracked per claim and factored into the downstream safety argument.

KPI Analyzers

Once the simulation completes, we must analyze the resulting logs to understand which safety requirement-relevant behaviors actually occurred. Halos SEF pairs simulation with a suite of automated KPI analyzers that consume the simulator's output alongside ground truth and validate adherence to the safety properties in question. For our L2++ intersection goal, the analyzer checks each simulated intersection traversal for collision events of Severity 2 or greater. For the AEB ghost-brake claim, it detects unwarranted trigger events in nominal driving. Some of our more advanced analyzers encode formally verifiable safety properties within them, such as HJ-Reachability based Safety Zones that NVIDIA has pioneered to understand which obstacles are actually safety-relevant to a given driving task.

A key architectural property of Halos SEF is the abstraction layer that sits between the simulator, the system under test (SUT), and the analyzer. Each of these three components speaks to the others through a standardized interface, which means any one of them can be swapped without disturbing the rest of the pipeline. We can exchange an open-loop replay simulator for a NuRec-backed closed-loop run, migrate from one version of the SUT to the next, or substitute a human-labeled ground truth source for an auto-labeled one, and the downstream analyzer continues to operate unchanged.

It is worth noting that safety analyzers are typically tuned to err on the side of caution, trading precision for recall. A missed failure is far more costly than a false alarm, and thus a layer of manual review is applied to triage flagged events and separate true failures from benign triggers. This reviewed output is what ultimately feeds the KPI numbers reported to Process Governance.

Process Governance: KPI Analysis and Reporting

Once the evidence supporting our safety case has been generated, we find ourselves once again in the Process Governance section, where we analyze and risk-assess the results.

Figure 8: We return to Halos Safety Evaluation Framework Simulation Core.

A raw failure rate is not, on its own, a safety argument. Results must be contextualized against their confidence bands, their data distributions, and the known precision and recall characteristics of the tools that produced them. They must be shown to be free from Risk Transfer[3] – situations where failure rates are within specifications, but their distributions are concentrated in unexpected ways.

Our result visualization tools and underlying data science platform are purpose-built to support this analysis, allowing safety engineers to drill from an aggregate KPI down into the individual driving logs, labels, and simulator outputs that contributed to it. This is also where NVIDIA-pioneered statistical tools such as Sim2Val come into play, allowing for efficient quantification of simulation realism in our resulting safety assurance.

Every piece of safety evidence that Halos SEF generates is designed to be traced back to a concrete, auditable requirement using our KPI management tools and result visualization platforms. This bookends the overall safety framework by connecting our KPI result analysis and risk assessment back to our requirements management.

The AV Safety Flywheel

As analysis of evaluation results feed back into our requirements management process, a feedback loop is formed, which enables us to iterate, and close performance gaps based on data-driven decision-making.

[3] It is not sufficient to develop an autonomous vehicle with an overall fatality rate lower than that of human drivers. We must also demonstrate that no group of individuals are disproportionately put at higher risk by this system. This concept is well explored in Chapter 5.2 of “How Safe Is Safe Enough” by Philip Koopman, where an example is given of a system that has a higher aggregate safety level than that of humans, but to which most accidents occur with pedestrians in high-visibility vests. Such a system may not be deemed acceptable to deploy.

Observation	Implication in next evaluation iteration
Insufficient data covering rain scenarios	Need to collect, mine, or augment more rainy data.
Insufficient sim realism	Need to improve simulator, or rely on an alternative
Safety KPIs below thresholds in minor-major intersections	Need to improve product in these areas, or limit the operational scope of the product, such as requesting driver takeover in these situations.
All KPIs being met across current ODDs	Opportunity to expand to new geographic regions. Update data distribution and evaluate readiness.

Table 2: Examples of data-driven decisions that Halos SEF enables through consecutive iterations of its safety flywheel.

As the product lifecycle progresses, the flywheel is turned until we establish a solid foundation of evidence to support the automotive safety case.

Making Halos Safety Evaluation Framework Credible and Auditable

In order to truly support an automotive safety case, all elements of the framework – from the datasets, the simulation technologies, to the KPI analyzers – need to be credible for the specific safety claim they support. UN-R ADS clause 7.2.1 and UN-R171 DCAS Annex 5 provide a useful structure for this credibility argument.

For researchers, this credibility case maps directly to the concept of a reproducible evaluation methodology: the same principle of making evaluation pipelines transparent, auditable, and replicable that underpins benchmark design in the computer vision community.

In Halos SEF, this credibility case is organized around five evidence areas:

- **Foundation and scope:** define the intended use of virtual testing, the relevant safety requirement, the test objective, ODD and system boundary, assumptions, limitations, fidelity expectations, validation strategy, acceptance criteria, and the criticality of potential toolchain errors for the safety argument.
- **Process and governance:** manage the data, people, and releases behind the evidence. This includes input and output data pedigree, data quality and coverage, traceability from KPI results back to scenarios, parameters, toolchain configuration and release, personnel competency, third-party inputs, and lifecycle support.
- **Technical verification:** demonstrate that the toolchain is implemented correctly and behaves robustly for valid inputs. Evidence includes code verification, parameter-space exploration, sanity and consistency checks, and bounded numerical errors such as discretization, rounding, and convergence effects.
- **Validation and uncertainty analysis:** show that simulation outputs correlate with relevant physical or real-world reference data for the intended use. This includes validation metrics, goodness-of-fit criteria, calibration, sensitivity analysis, uncertainty quantification, and safety margins where uncertainty affects interpretation of results.
- **Final suitability assessment:** synthesize the evidence and state whether the toolchain and SEF element are suitable for the defined claim, scope, release, assumptions, and limitations.

Halos SEF is intended to make this evidence structure explicit through an overall SEF Credibility Manual and supporting manuals for individual SEF elements. These manuals map the credibility expectations above to evidence generated through established standards and practices, including ISO 26262, ISO 21448, ISO/PAS 8800, ISO 34502, ISO 34503, and ISO/IEC 17025.

Conclusion & Outlook

The shift toward AI-based architectures is changing what it means to demonstrate AV safety. Traditional functional safety methods, built around deterministic software and deductive failure analysis, are necessary but no longer sufficient. AI-driven systems demand a data-driven, statistical treatment of safety – and that treatment, in turn, demands an evaluation framework capable of generating evidence at a scale and fidelity that the industry has not previously had to contend with.

Halos SEF is NVIDIA's answer to that challenge. By bringing together a Data Foundation, a Simulation Core, and Process Governance into a single, coherent recipe, we provide a specific and actionable path from safety requirements to a reproducible, standards-compliant safety case. We invite our partners and customers to build this future with us, so that the promise of AI-driven autonomy can be delivered with the statistical rigor that the safety of the driving public demands.

For researchers, Halos SEF represents a broader principle that extends beyond autonomous vehicles: a physical AI system, from robots to surgical assistants to smart infrastructure – moves from research labs into the real world, the same rigorous, reproducible evaluation methodology will be essential. The neural rendering, generative simulation, and statistical analysis tools developed for AV safety evaluation are foundational building blocks for the entire physical AI research agenda that NVIDIA is advancing with the community.