Defining Operational Conditions for Safety-Critical AI-Based Systems from Data

Johann M. Christensen, Elena Hoemann, Frank Köster, Sven Hallerbach

Institute for AI Safety and Security, German Aerospace Center (DLR)

1 Who Defines the ODD?

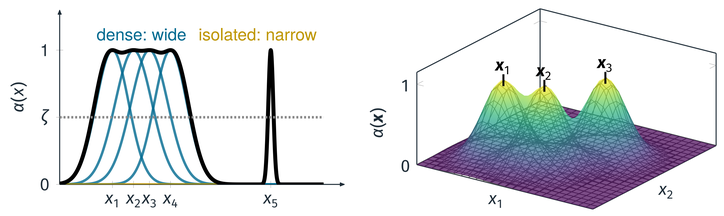
Certification builds on the operational design domain

The ODD specifies the conditions under which an AI-based system is intended to operate safely [1–3].

- **Experts** define it a priori, which fails without knowledge of the system, for systems too complex to understand, or when parameters interact in ways nobody can specify beforehand
- **Polytopes and convex hulls** [4] enclose voids in which no data exist
- **Prior data-driven ODDs** [5] are neither order-independent nor comparable

***auto*SAFE** derives the ODD *a posteriori* from recorded data: every sample becomes the anchor of a kernel, without expert knowledge.

2 Every Sample Becomes a Kernel



Each anchor $\mathbf{x}_i$ carries an RBF kernel, their noisy-OR [6] (black, surface) is the affinity:

$$\alpha_i(\mathbf{x}) = \exp\left(-\tfrac{1}{2}(\mathbf{x}-\mathbf{x}_i)^\top \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\mathbf{x}_i)\right) \tag{1}$$

$$\alpha(\mathbf{x}) = 1 - \prod_i \left(1-\alpha_i(\mathbf{x})\right) \in [0,1] \tag{2}$$

Log-margin rule. Products of many $1-\alpha_i$ underflow. The log-survival turns the product into a stable sum [7]:

$$S(\mathbf{x}) = \ln\left(1-\alpha(\mathbf{x})\right) = \textstyle\sum_i \ln\left(1-\alpha_i(\mathbf{x})\right) \tag{3}$$

$$\mathbf{x} \in \mathcal{O} \iff \alpha(\mathbf{x}) \geq \zeta \iff S(\mathbf{x}) \leq \ln(1-\zeta) \tag{4}$$

3 The *auto*SAFE Algorithm

Input: in-distribution (ID) records, out-of-distribution (OOD) samples, OOD bound ξ, false-exclusion budget ε

1. **De-duplicate the records:** one anchor per acquisition-resolution cell, in canonical order ⇒ deterministic and permutation-stable
2. **Calibrate the kernel widths** from the nearest-neighbor gaps d^*_{ik}, in units of the median gap $\tilde{d}$:

$$\sigma^{(i)}_{kk} = (\kappa-\lambda)\exp\left(-\eta d^*_{ik}\right) + \lambda \tag{5}$$

$$\kappa = (s_N \tilde{d})^2, \quad \eta = \gamma/\tilde{d}, \quad \lambda = \lambda_{\text{rel}}\kappa$$

   Dense anchors get wide kernels, isolated ones narrow kernels; the relative floor λ_{rel} bounds the condition number ⇒ unit-invariant
3. **Enforce the OOD bound:** while $\max_{\mathbf{x}\in\mathcal{D}_{\text{OOD}}} \alpha(\mathbf{x}) > \xi$, shrink the dominant kernel, $\boldsymbol{\Sigma}_{i^*} \leftarrow c\,\boldsymbol{\Sigma}_{i^*}$ ⇒ provably terminates
4. **Calibrate the threshold** split-conformally [8] on held-out records, with their $S(\mathbf{x})$ sorted in descending order:

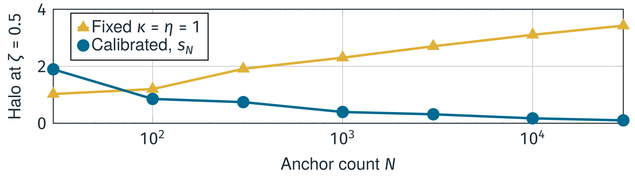
$$\hat{\zeta} = 1 - \exp\left(S_{(k)}\right), \quad k = \lfloor \varepsilon(N_{\text{cal}}+1) \rfloor$$

$$\mathbb{P}\left(\alpha(\mathbf{x}) < \hat{\zeta}\right) \leq \varepsilon \tag{6}$$

Output: affinity α and threshold $\hat{\zeta}$: a deterministic, bounded ODD that doubles as a runtime monitor

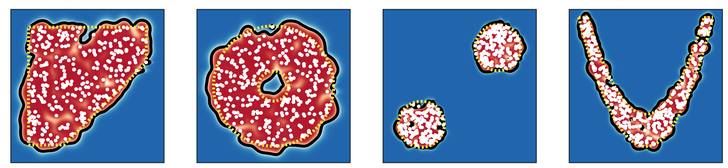
4 More Data, Tighter ODD

With s_N, the ODD provably converges in volume from both sides.



5 Synthetic Benchmarks

As accurate as density estimators, plus the guarantees



AUPR, N = 1000	Linear	Annulus	Two blobs	Banana
***auto*SAFE**	0.992	0.990	0.985	0.970
KDE [9]	0.996	0.995	0.996	**0.986**
GMM (BIC) [10]	**0.998**	**0.998**	**0.997**	0.983
k-NN distance	0.991	0.991	0.984	0.970
OC-SVM [11]	0.995	0.817	**0.997**	0.337
Isolation Forest [12]	0.952	0.918	0.972	0.821
Convex hull	0.988	0.901	0.482	0.332

Within 0.016 of the best AUPR: the gain is the assurance-oriented construction, not accuracy. The convex hull fails on non-convex ODDs.

6 An Aviation Use Case

Accepted points in a gap of the data, lower is better

Aircraft collision-avoidance states [13], one region held out:

54.7 %	4.2 %	0.55 %
Convex hull	*auto*SAFE	*auto*SAFE + OOD

Against the known ODDs of the system:

> 0.99	0.46–0.59	0.00997
Precision, all ζ	Recall at $\hat{\zeta}_{0.1}$	OOD affinity, ξ = 0.01

Take-home

1. An ODD can be derived **from data alone**: no expert-drawn boundaries, no convex-hull voids.
2. **Deterministic, bounded, unit-invariant**, with proven OOD exclusion and a conformal false-exclusion guarantee.
3. It shows where data **provide support**, not which conditions are safe: one evidence artifact in a safety case.

[1] European Union Aviation Safety Agency (EASA). EASA Concept Paper: Guidance for Level 1 & 2 Machine Learning Applications. 2024.
[2] International Organization for Standardization (ISO). Road vehicles — Test scenarios for automated driving systems — Specification for operational design domain. 2023.
[3] SAE International. Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles: J3016. 2021.
[4] V. Nenchev. One Stack, Diverse Vehicles: Checking Safe Portability of Automated Driving Software. 2025.
[5] F. Kaakai et al. Data-Centric Operational Design Domain Characterization for Machine Learning-Based Aeronautical Products. 2023.
[6] J. Pearl. Probabilistic Reasoning in Intelligent Systems. 2014.
[7] M. Mächler. Accurately Computing log(1− exp(− |a|)) Assessed by the Rmpfr package. 2015.
[8] A. N. Angelopoulos et al. Conformal Prediction: A Gentle Introduction.
[9] L. Devroye et al. Detection of Abnormal Behavior Via Nonparametric Estimation of the Support.
[10] G. J. McLachlan et al. Finite Mixture Models. 2000.
[11] B. Schölkopf et al. Estimating the Support of a High-Dimensional Distribution.
[12] F. T. Liu et al. Isolation Forest. 2008.
[13] K. D. Julian et al. Guaranteeing Safety for Neural Network-Based Aircraft Collision Avoidance Systems. 2019.