Unifying Runtime Monitoring Approaches for Safety-Critical Machine Learning: Application to Vision-Based Landing

Mathieu DARIO
PhD at LAAS-CNRS / Thales
Toulouse, France

Under the supervision of
F. Chenevier, K. Delmas, J. Guerin, J. Guiochet

ICPR 2026
International Conference on Pattern Recognition LYON

Our paper

19/08/2026

mathieu.dario@laas.fr
mathieu.dario@thalesgroup.com

Context & Use Case

AI/ML in Aeronautics

- **Safety-Critical** applications
- **Perception tasks** (computer vision)
- **Deep neural networks** (DNNs)

Context & Use Case

AI/ML in Aeronautics

- **Safety-Critical** applications
- **Perception tasks** (computer vision)
- **Deep neural networks** (DNNs)

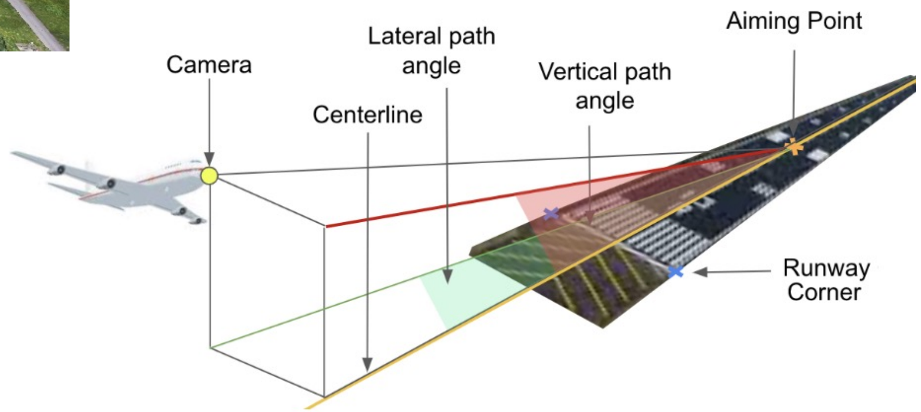


Figure. Geometry of a landing [1].

[1] C. Cappi *et al.*, 'How to design a dataset compliant with an ML-based system ODD?', in *ERTS 2024*

The Issue with Deep Learning

A strictly regulated field

- Deep Networks
  - ❌ Adverse / Unforeseen inputs
  - ❌ Black-box
  - ❌ Overconfidence when incorrect

➔ Reliability, integrity required

➔ Certification [EASA]

The Issue with Deep Learning

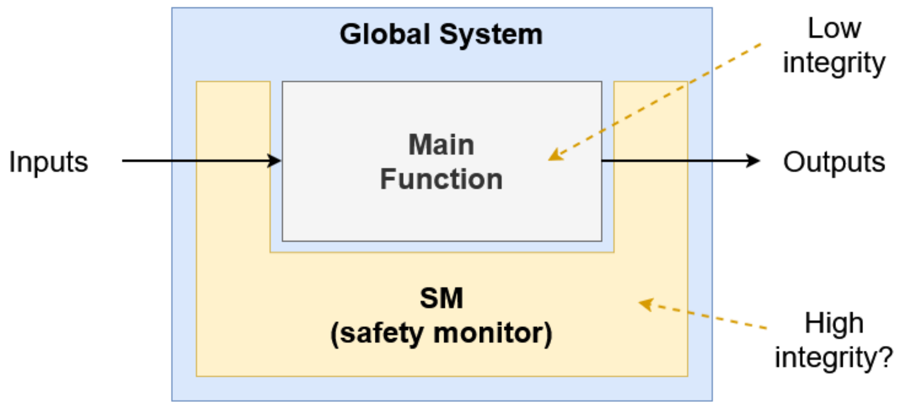
A strictly regulated field

- Deep Networks
  - ❌ Adverse / Unforeseen inputs
  - ❌ Black-box
  - ❌ Overconfidence when incorrect

→ Reliability, integrity required

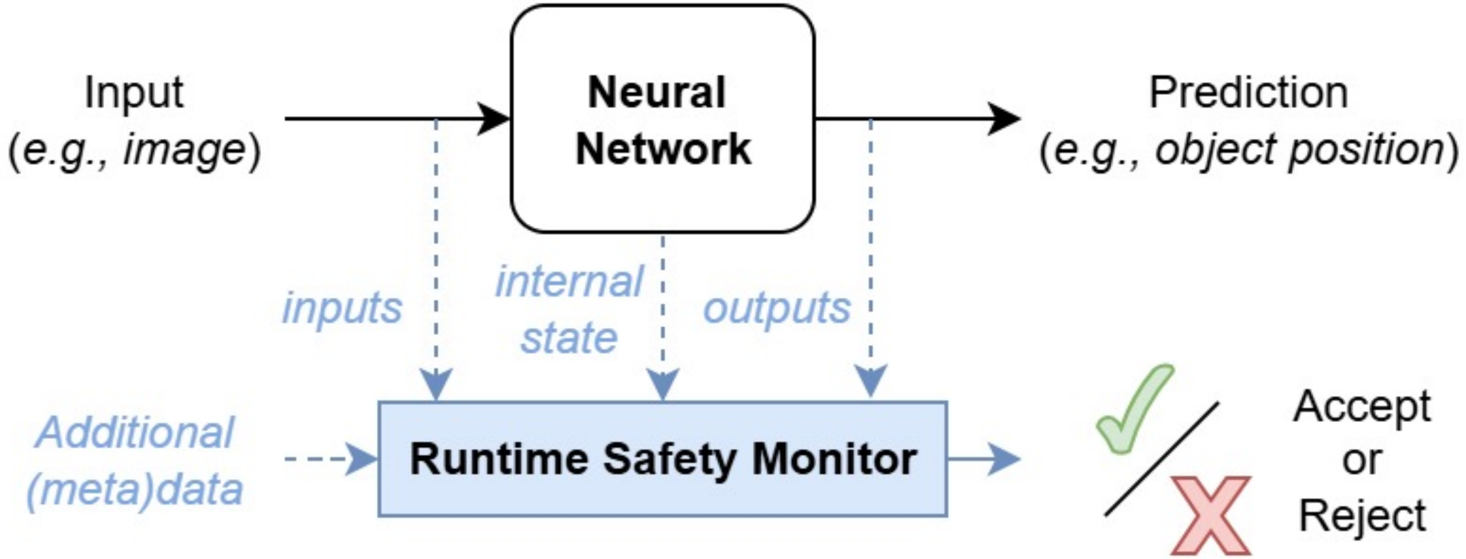
→ Certification [EASA]

Runtime Safety Monitors



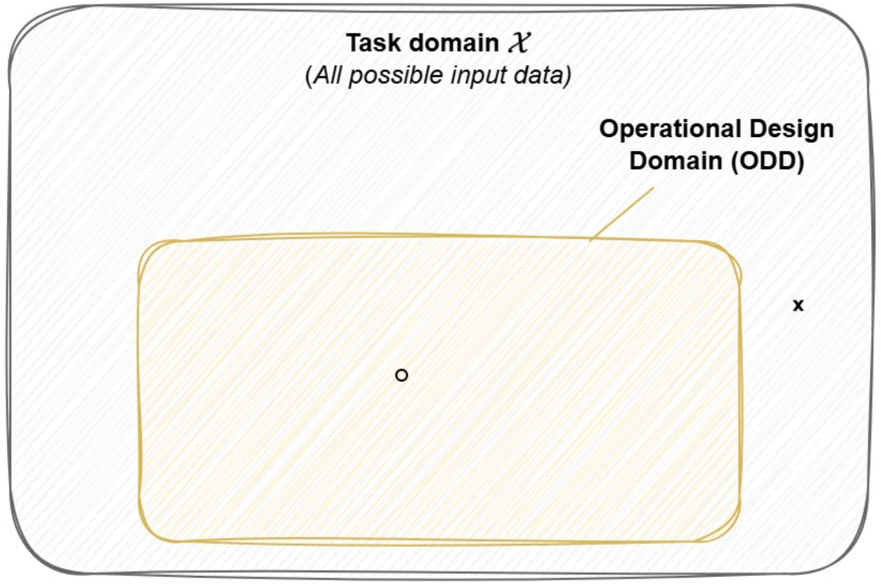
- From the *dependability community*

Safety Monitors for Machine Learning Components (DNNs)



In aeronautics safety standards

ODD (Operational Design Domain)



o ML prediction is correct

x ML prediction is wrong

Operational Design Domain

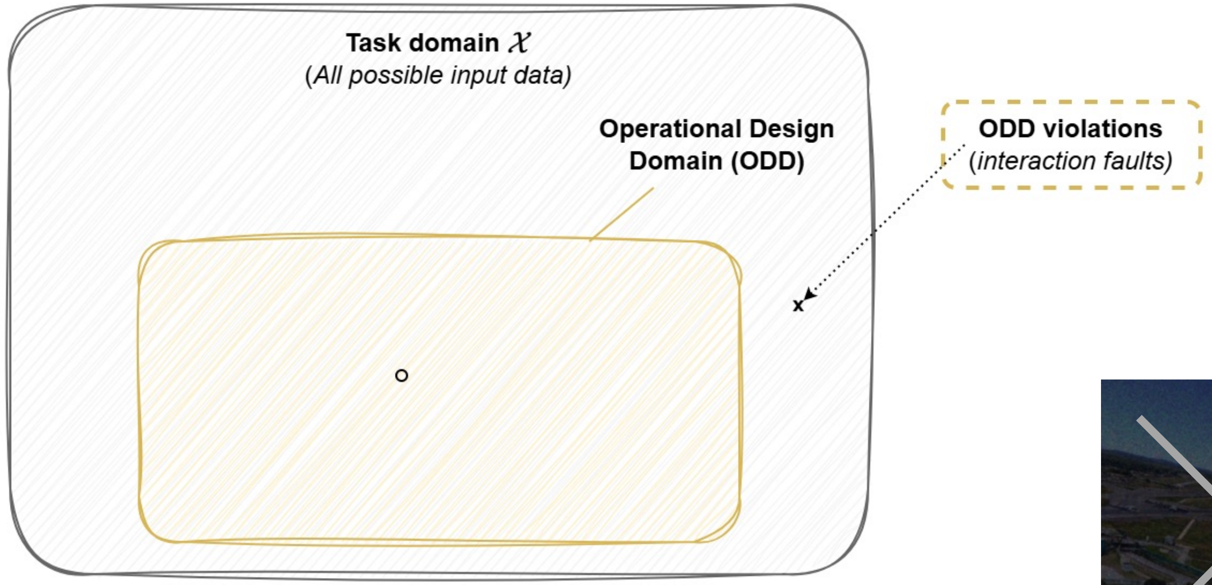
"***Operating conditions*** *under which a given AI/ML constituent is specifically* ***designed*** *to function as intended, including but not limited to environmental, geographical, and/or time-of-day."* EASA [2]

[2] EASA, 'EASA concept paper: Guidance for level 1 & 2 machine learning applications issue 02', Mar. 2024.

ODD Violations

Operational Design Domain

*"**Operating conditions** under which a given AI/ML constituent is specifically **designed** to function as intended, including but not limited to environmental, geographical, and/or time-of-day."* EASA [2]



o ML prediction is correct

x ML prediction is wrong

Table. Parameters and ranges for the Generic Landing Approach Cone [1].

Parameter	range
Along track distance	[0.08, 3] NM
Vertical path angle	[-2.2, -3.8]°
Lateral path angle	[- 4, 4] °
Yaw	[-10,10] °
Pitch	[-8,0] °
Roll	[-10,10] °

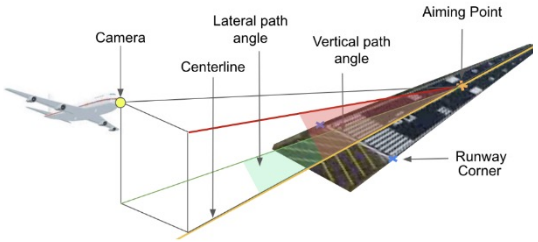
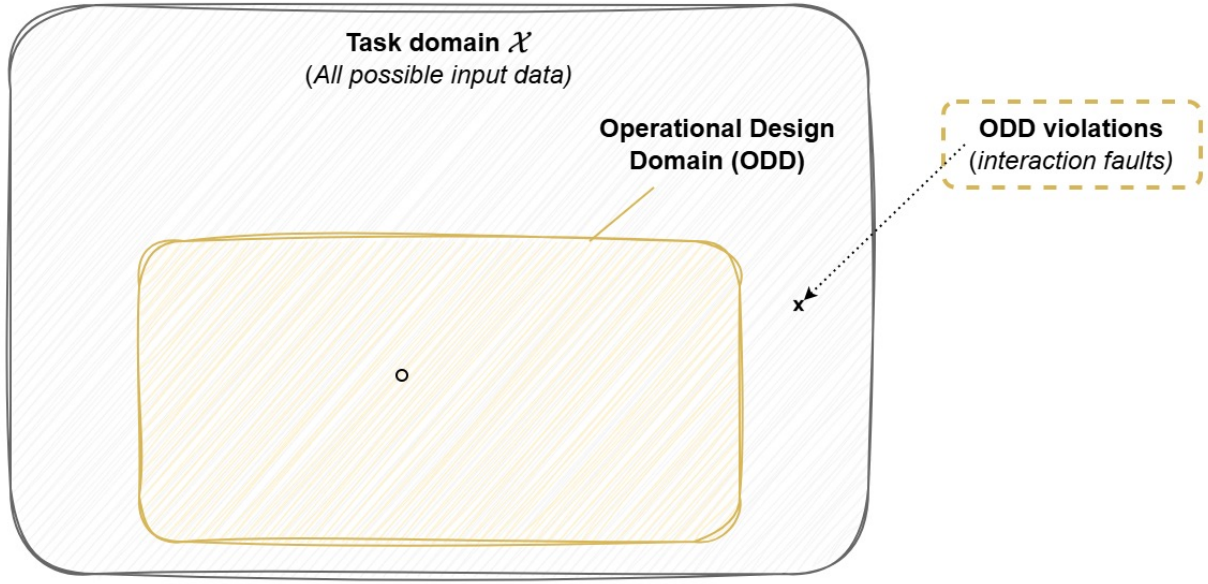


Figure. Geometry of a landing as described in [1].

[2] EASA, 'EASA concept paper: Guidance for level 1 & 2 machine learning applications issue 02', Mar. 2024.
[1] C. Cappi *et al.*, 'How to design a dataset compliant with an ML-based system ODD?', in *ERTS 2024*

Monitoring the ODD



o ML prediction is correct

x ML prediction is wrong

Operational Design Domain

*"**Operating conditions** under which a given AI/ML constituent is specifically **designed** to function as intended, including but not limited to environmental, geographical, and/or time-of-day."* EASA [2]

- **Description.**
  ODD parameters + ranges, constraints.

  → At *system-level* (system variables) [1,3]
  → At *image-level* / ML-level [1].

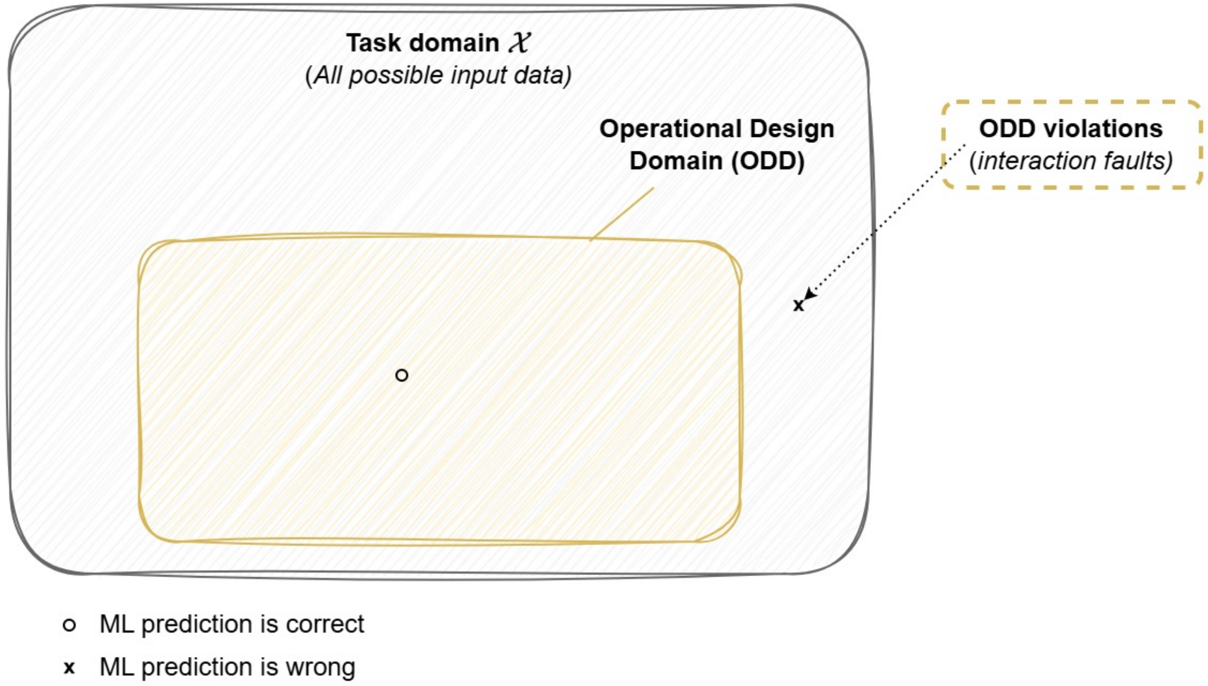
- **Monitoring.**
  - <u>Rule-based</u> approaches (high integrity)
  - <u>Data-driven</u> approaches (low integrity)

[2] EASA, 'EASA concept paper: Guidance for level 1 & 2 machine learning applications issue 02', Mar. 2024.
[1] C. Cappi *et al.*, 'How to design a dataset compliant with an ML-based system ODD?', in *ERTS 2024*
[3] C. Torens *et al.*, 'Runtime monitoring of operational design domain to safeguard machine learning components', *CEAS Aeronaut J*, 2025.

Monitoring the ODD

Operational Design Domain

*"**Operating conditions** under which a given AI/ML constituent is specifically **designed** to function as intended, including but not limited to environmental, geographical, and/or time-of-day."* EASA [2]



- **Description.**
  ODD parameters + ranges, constraints.

  → At *system-level* (system variables) [1,3]
  → At *image-level* / ML-level [1].

- **Monitoring.**
  - Rule-based approaches (high integrity)
  - Data-driven approaches (low integrity)

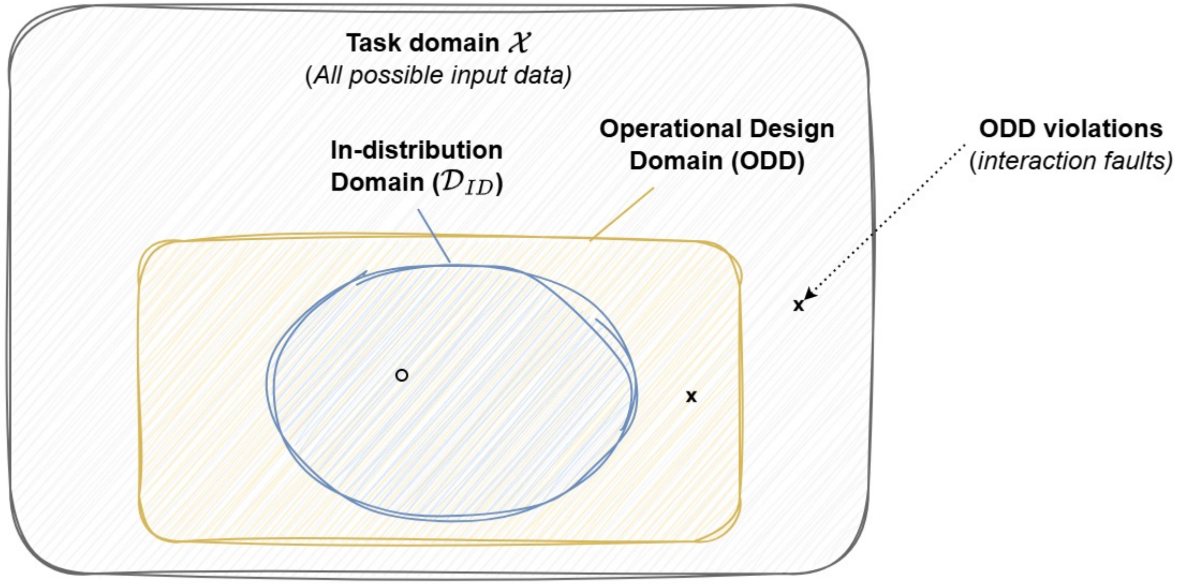
Parameter	range
Along track distance	[0.08, 3] NM
Vertical path angle	[-2.2, -3.8]°
Lateral path angle	[- 4, 4] °
Yaw	[-10,10] °
Pitch	[-8,0] °
Roll	[-10,10] °

[2] EASA, 'EASA concept paper: Guidance for level 1 & 2 machine learning applications issue 02', Mar. 2024.
[1] C. Cappi *et al.*, 'How to design a dataset compliant with an ML-based system ODD?', in *ERTS 2024*
[3] C. Torens *et al.*, 'Runtime monitoring of operational design domain to safeguard machine learning components', *CEAS Aeronaut J*, 2025.

In the ML community

In-Distribution Domain

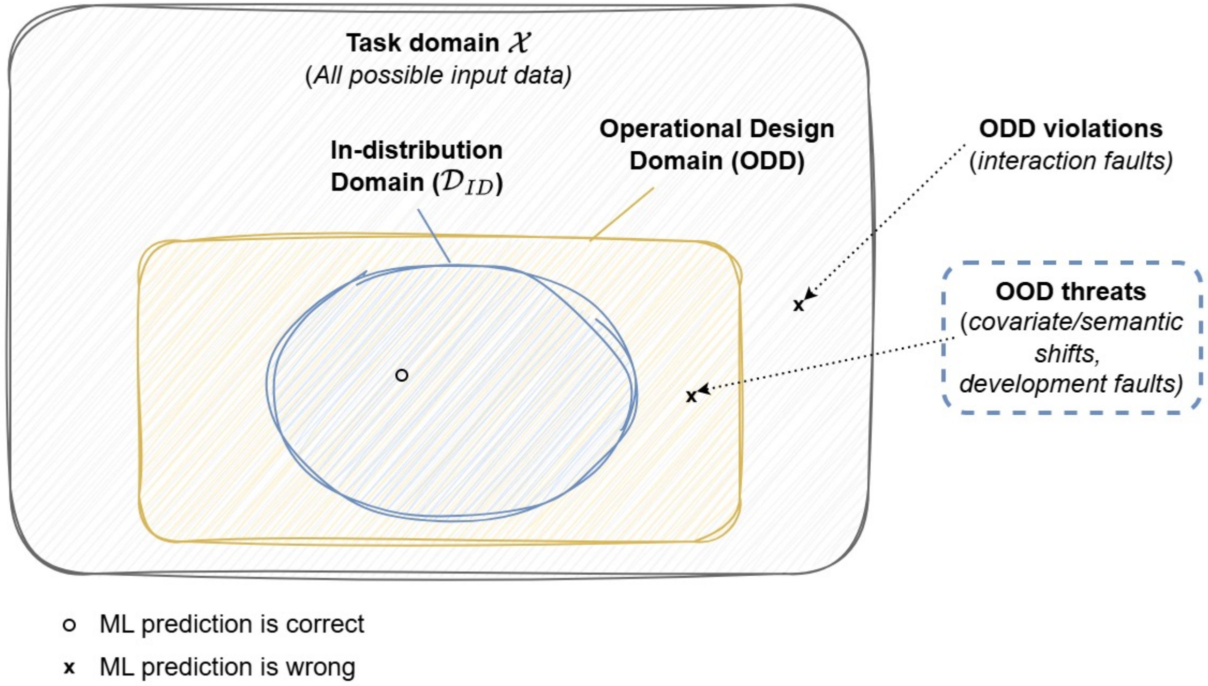
In-distribution domain = data from same distribution than *training* data



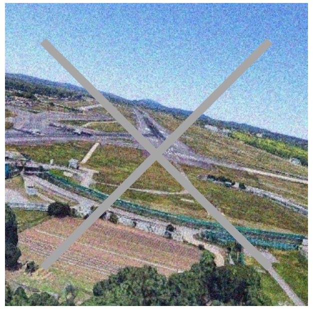
- o ML prediction is correct
- x ML prediction is wrong

OOD (Out-of-Distribution) Threats

In-distribution domain = data from same distribution than training data



Overexposure effect

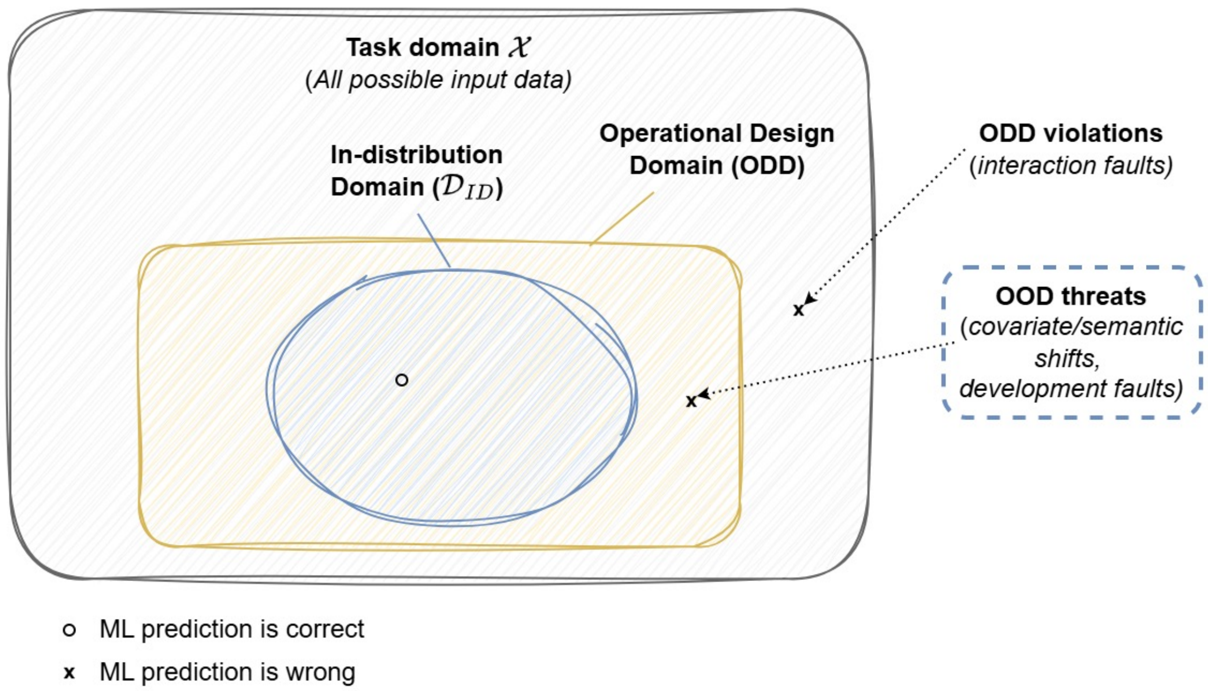


Gaussian noise

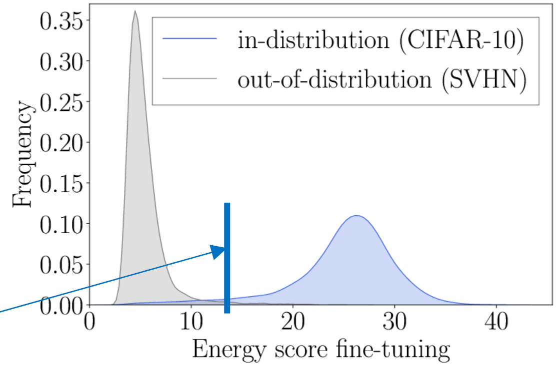
Covariate / *Semantic* shifts.

OOD Detection

In-distribution domain = data from same distribution than training data



OOD score threshold

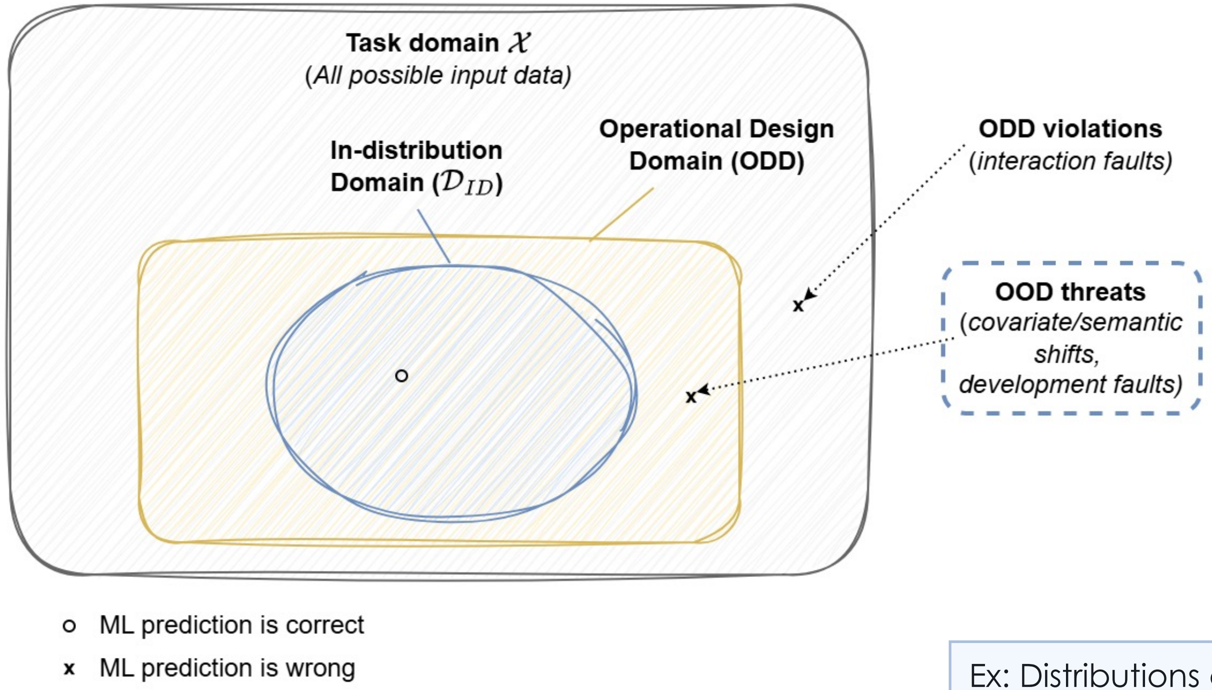


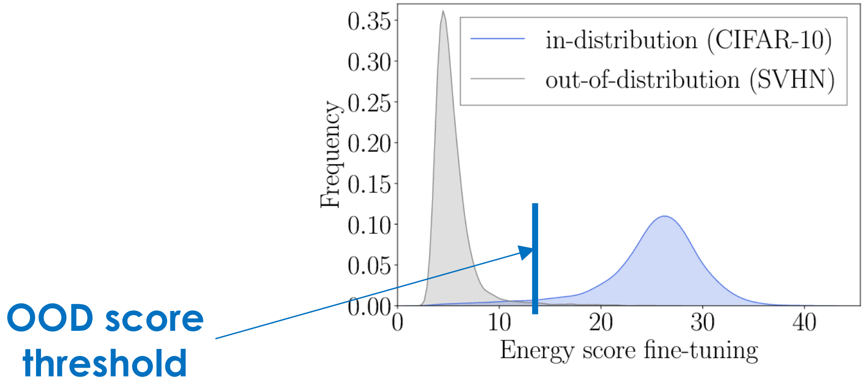
[Energy-based Out-of-Distribution Detection, Liu et al. 2020]

- Density-based
- Proximity-based
- Reconstruction-based

OOD Detection

In-distribution domain = data from same distribution than training data





[Energy-based Out-of-Distribution Detection, Liu et al. 2020]

Ex: Distributions of images meta-properties [3]:

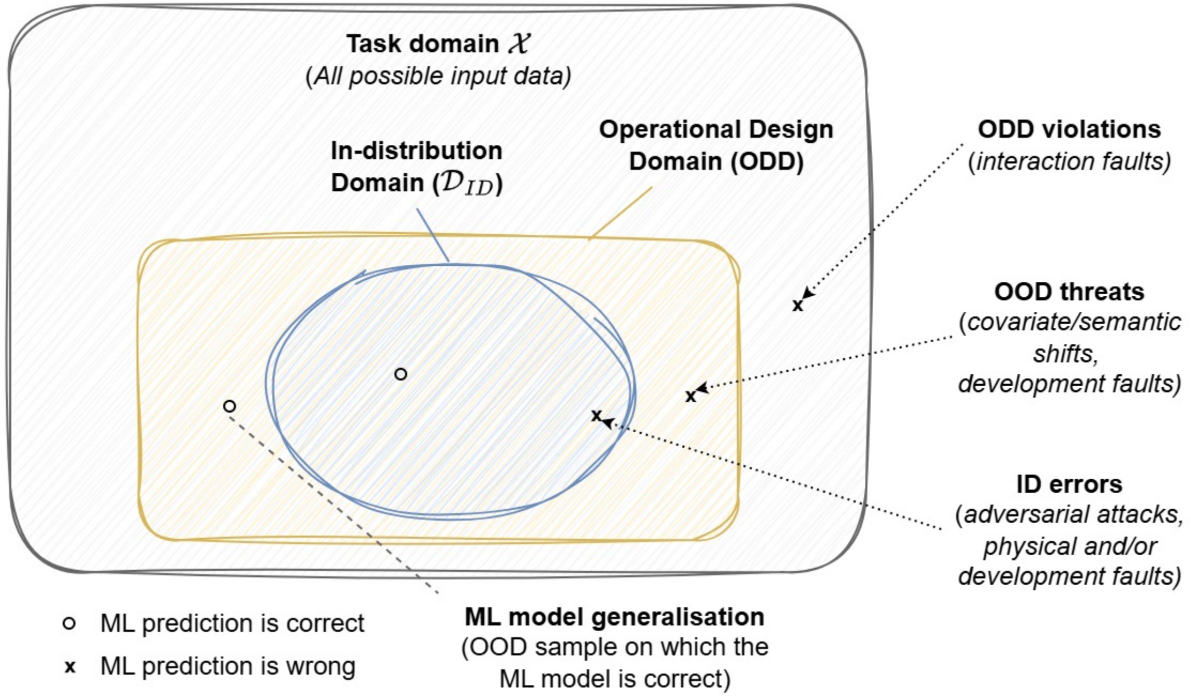
(**brightness**, **saturation**, **entropy**, **edge amount**)

- Density-based
- Proximity-based
- Reconstruction-based

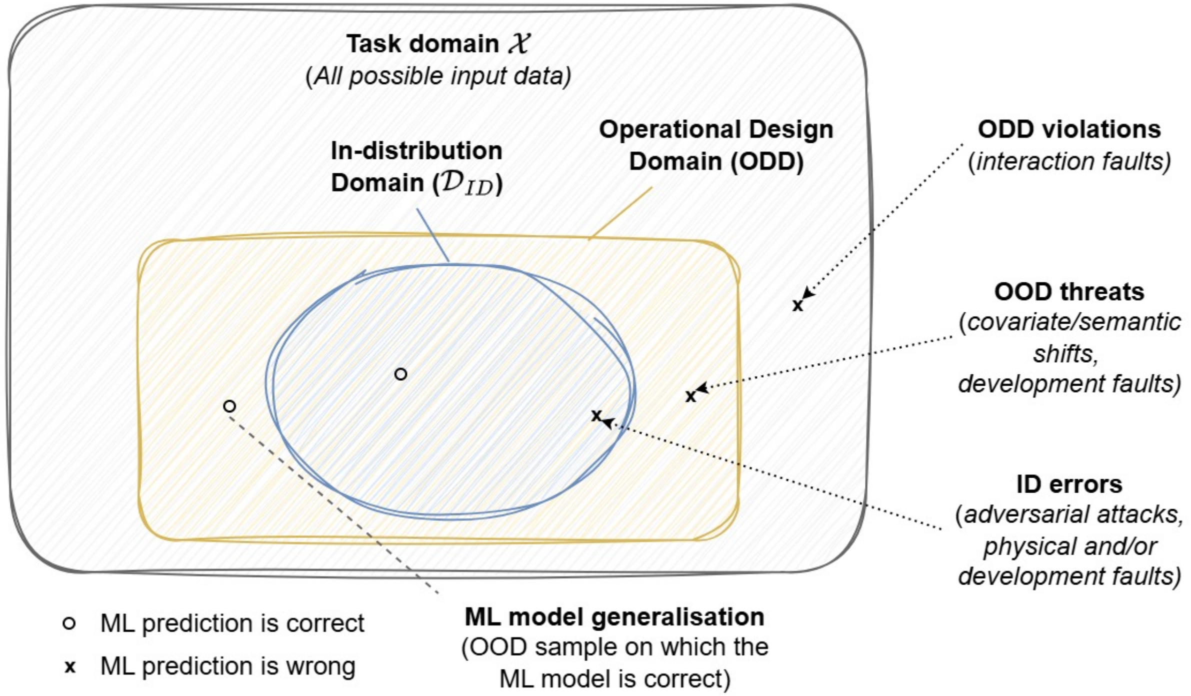
[3] C. Torens *et al.*, 'Runtime monitoring of operational design domain to safeguard machine learning components', *CEAS Aeronaut J*, Jul. 2025.

New paradigm from ML dependability

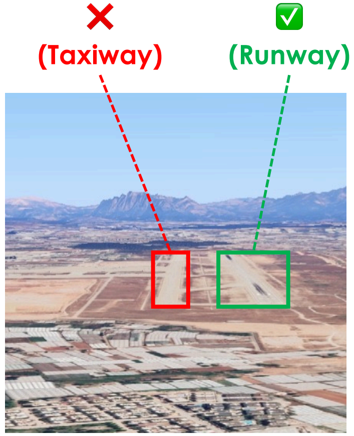
The Misalignment (In-Distribution Errors)



The Misalignment (In-Distribution Errors)

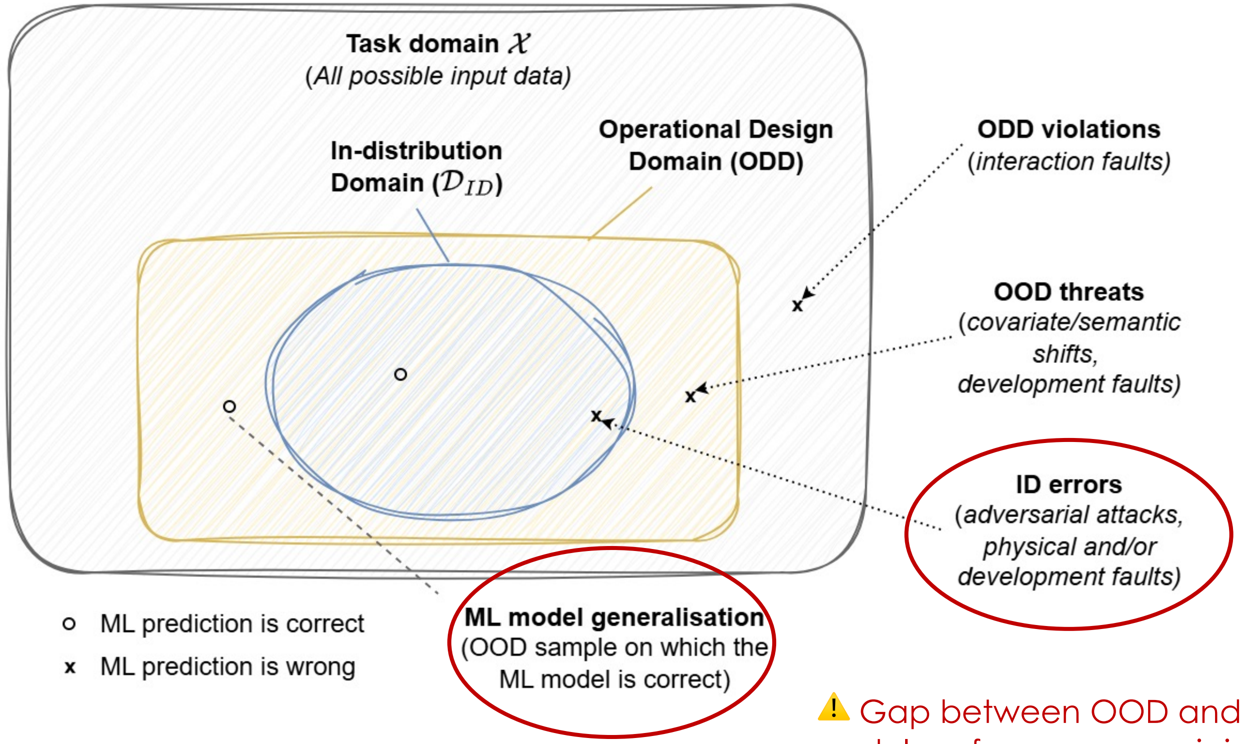


Adversarial attack



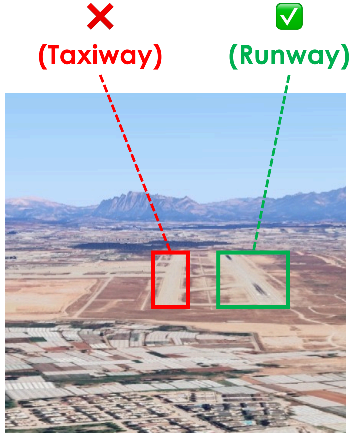
Complex scenario

Justifying the Model Scope Concept



⚠ Gap between OOD and model performance… origin of **Model Scope** concept [4]

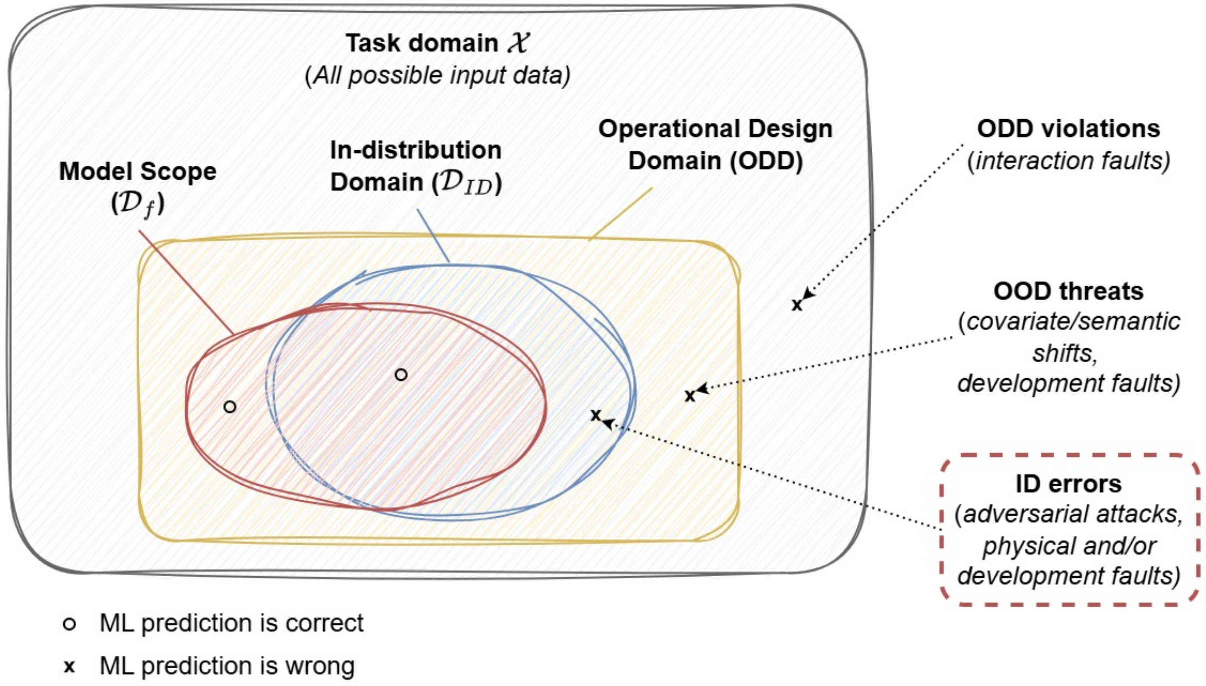
Adversarial attack



Complex scenario

[4] Guérin, J. Out-of-distribution detection is not all you need. In *Proceedings of the AAAI conference on artificial intelligence*. Jun 2023.

OMS (Out-of-Model-Scope) Detection Task

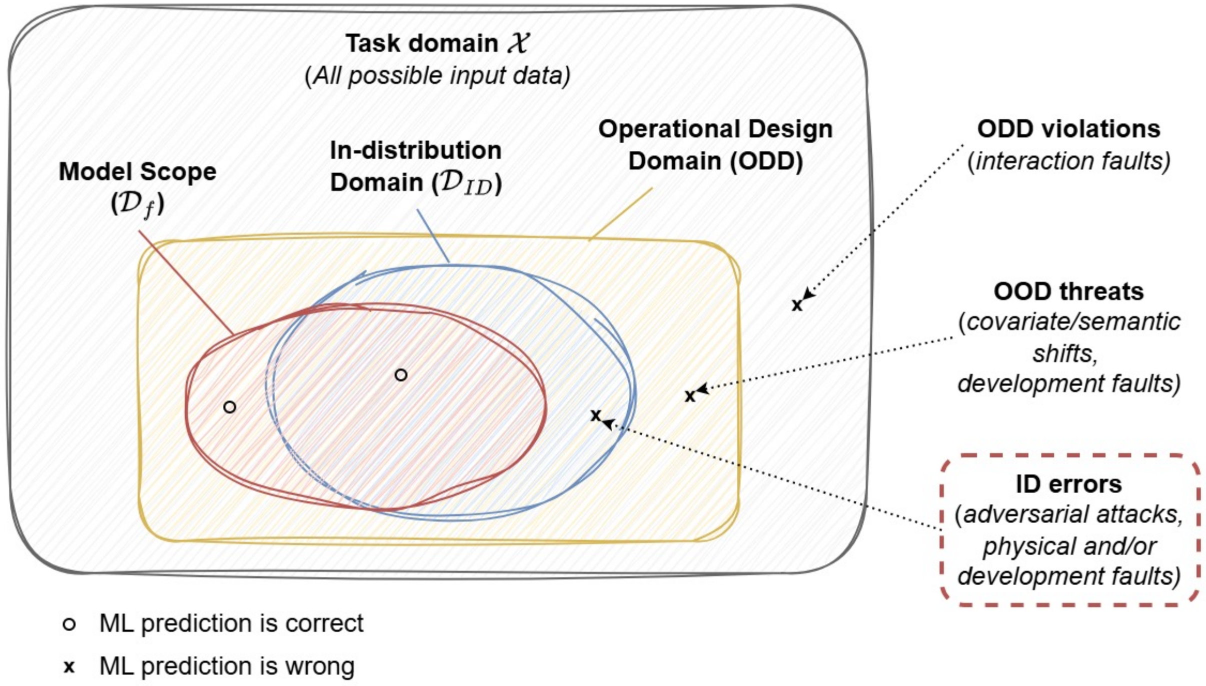


Model Scope = set of data instances where model is correct [4].

- Specific to one ML model/NN
- **Goal.** To detect model errors
- **How?** Using *anomaly detection* on outputs/internal behaviour as proxy

[4] Guérin, J. Out-of-distribution detection is not all you need. In *Proceedings of the AAAI conference on artificial intelligence*. Jun 2023.

OMS Subclassification

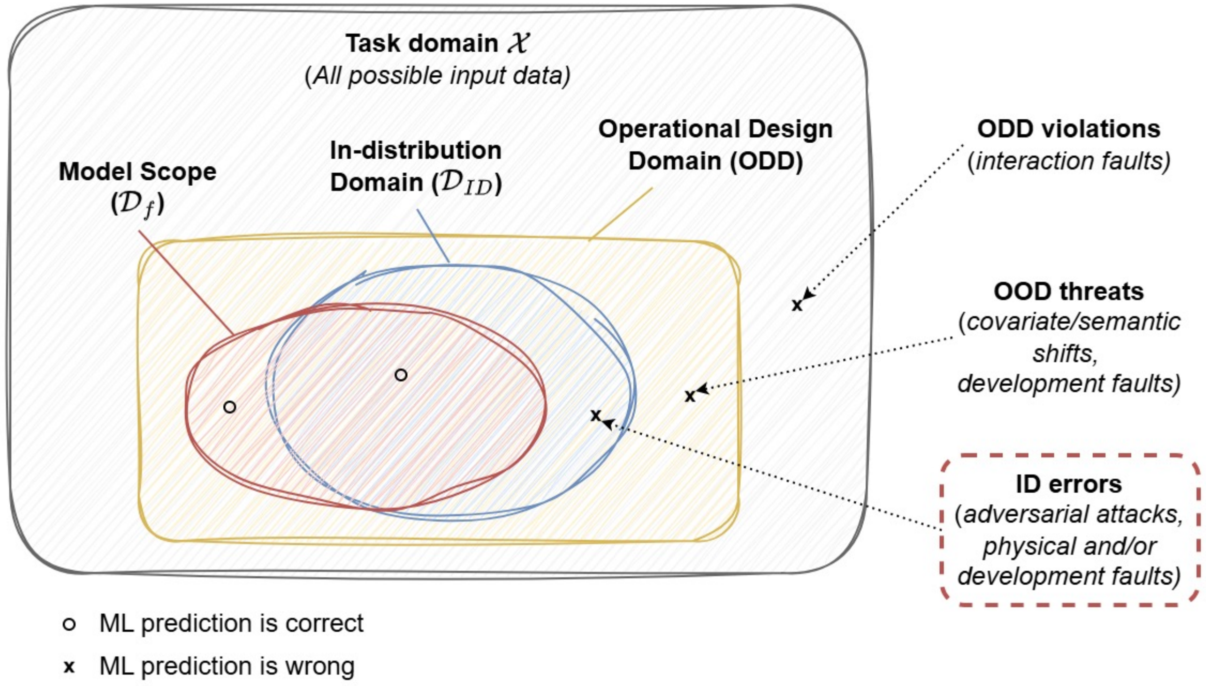


Proposed subclassification of OMS detectors

OMS detection methods

- **Feature-based** (e.g., activation patterns, outside-the-box, Mahalanobis, …)
- **Uncertainty-based** (e.g., MCDropout, Ensembles, Conformal Predictions, …)
- **Confidence-score** (e.g., threshold on softmax, energy score, logits score, …)
- **Consistency-based** (e.g., statistics on predictions, temporal/spatial consistency, …)

OMS Subclassification



Proposed subclassification of OMS detectors

OMS detection methods

- **Feature-based** (e.g., activation patterns, outside-the-box, Mahalanobis, …)
- **Uncertainty-based** (e.g., MCDropout, Ensembles, Conformal Predictions, …)
- **Confidence-score** (e.g., threshold on softmax, energy score, logits score, …)
- **Consistency-based** (e.g., statistics on predictions, temporal/spatial consistency, …)

Ex: **Box-Abstraction (BAM)** detector on the logits vector (classification head) [5].

[5] C. Wu, W. He, C.-H. Cheng, X. Huang, and S. Bensalem, 'BAM: Box Abstraction Monitors for Real-time OoD Detection in Object Detection', in *2024 IEEE/RSJ IROS*, Oct. 2024

A Unified & Structured Approach

Operational Design Domain (ODD) Monitoring

From the automotive & aeronautics fields

Is the system used in intended operating conditions?

[EASA, 2024], [Kaakai, 2023]
[Torens, 2023], [Torens, 2025]

Out-of-Distribution (OOD) Detection Task

From ML/DL community

Detect inputs statistically different from ML training data.

[Hendrycks, 2017], [Lee, 2018]
[Sun, 2022], [Liu, 2020]
[Henzinger, 2020], [He, 2024]

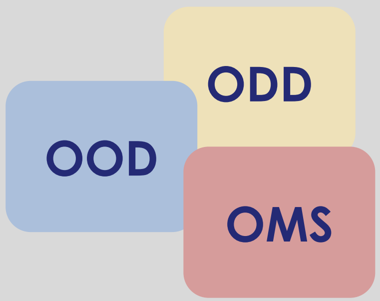
Out-of-Model-Scope (OMS) Detection Task

From ML Safety & Dependability community

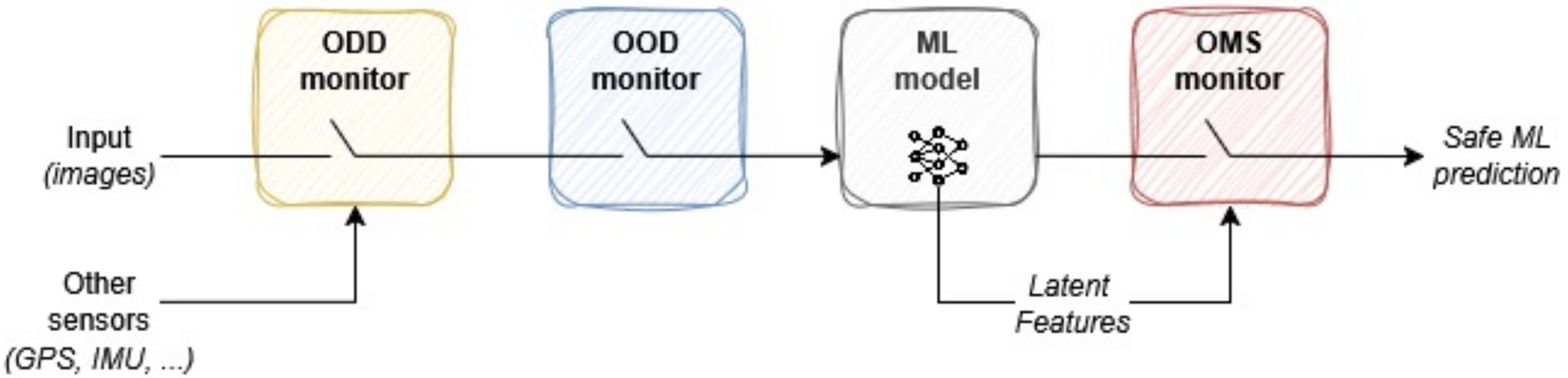
Detect inputs on which NN fails.
***Failure** detection task.*

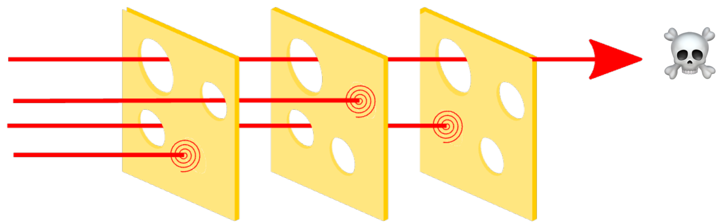
[Guerin, 2023], [Hashemi, 2024]
[Cheng, 2019]
[Wang, 2020], [Rahman, 2021]

Unified, structured, justified Monitoring Methodology

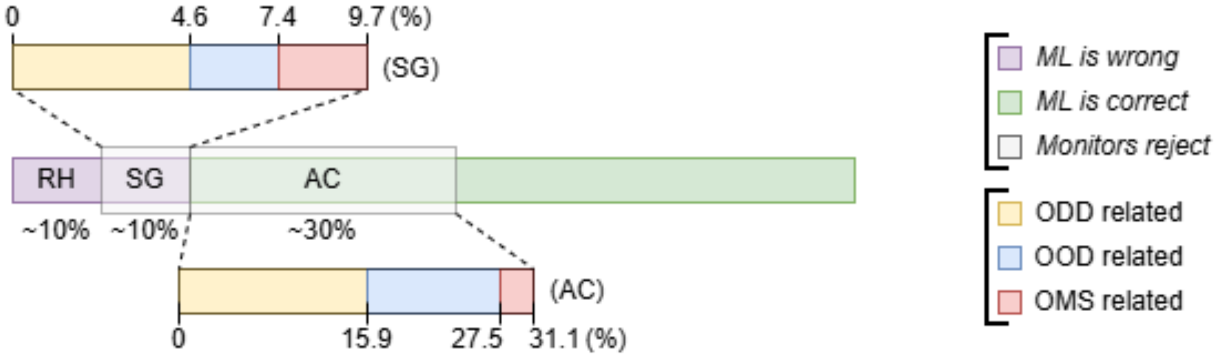


The Unified Sequential Architecture





Experimental Validation & Systemic Behavior



Safety Gain (SG), Residual Hazard (RH), and Availability Cost (AC) [6]
when using all monitors combined.

➔ **Complementarity of the monitoring families.**

➔ **Safety / Availability trade-off.**

[6] J. Guerin, R. S. Ferreira, K. Delmas, and J. Guiochet, 'Unifying Evaluation of Machine Learning Safety Monitors', in *2022 IEEE ISSRE*.

Takeaways & Perspectives

> Unified methodology for ML safety monitoring

- Conceptual bridge between ODD, OOD and OMS.
- Justified sequential monitoring funnel.
- Addressing the different threats to ML models.

> Perspectives.

- Experiment version 2 – more performant monitors and more complex runway detection use case
- Task-specific metrics for each monitoring class
- Formulating a safety assurance case to justify integrity of this monitoring approach

Unifying Runtime Monitoring Approaches for Safety-Critical Machine Learning: Application to Vision-Based Landing

Mathieu DARIO
PhD at LAAS-CNRS / Thales
Toulouse, France

Under the supervision of
F. Chenevier, K. Delmas, J. Guerin, J. Guiochet

International Conference on Pattern Recognition ICPR 2026 LYON

Our paper

19/08/2026

mathieu.dario@laas.fr
mathieu.dario@thalesgroup.com