NTNU

Norwegian University of Science and Technology

Marie Skłodowska-Curie Actions






# ENHANCHING TRUSTWORTHY AI INTEGRATION IN SAFETY-CRITICAL SYSTEM

**Dina Tri Utari**

Department of Mechanical and Industrial Engineering, NTNU

Main Supervisor: Prof. Shen Yin
Co-Supervisor: Prof. David Camacho (UPM), Prof. Volker Stolz (HVL), Prof. Francesco Piccialli (UNINA)

https://tuai.eu/

## Main Objective

To establish and validate a **comprehensive framework** for developing, evaluating, and deploying **trustworthy and sustainable AI framework** in safety-critical CPS.

The work focuses on enhancing transparency, fault tolerance, and robustness under real-world uncertainty while ensuring alignment with **industrial safety requirements.**

## Expected Results

1. Framework and metrics for assessing AI trustworthiness and sustainability.
2. Prototype algorithms and software tools for trustworthy AI integration.

NTNU | Norwegian University of Science and Technology

# Toward a Trustworthiness Index

## for AI-Enabled Safety-Critical CPS

*A Multi-Dimensional Framework*

**Dina Tri Utari** | Shen Yin
Department of Mechanical and Industrial Engineering
Norwegian University of Science and Technology (NTNU)

# MOTIVATION

## AI in Safety-Critical CPS

- Autonomous Vehicles
  Deep learning for perception, object detection, trajectory prediction
- Industrial Automation
  Predictive maintenance, quality inspection, collaborative robotics
- Medical Devices
  Disease diagnosis, treatment suggestions, equitable care
- Aviation
  Flight control, collision avoidance

## Real-World Failures

**2018 — Uber AV Fatal Crash**

Perception failure → pedestrian death. Demonstrates silent ML failure in safety-critical context.

**2020 — Tesla Autopilot**

Collision with stationary emergency vehicles — edge case beyond training distribution.

**2021 — Medical AI Bias**

Systematic underdiagnosis in Black patients — bias hidden by aggregate accuracy metrics.

**No unified, quantitative way exists to assess, compare, or certify AI trustworthiness across these domains.**

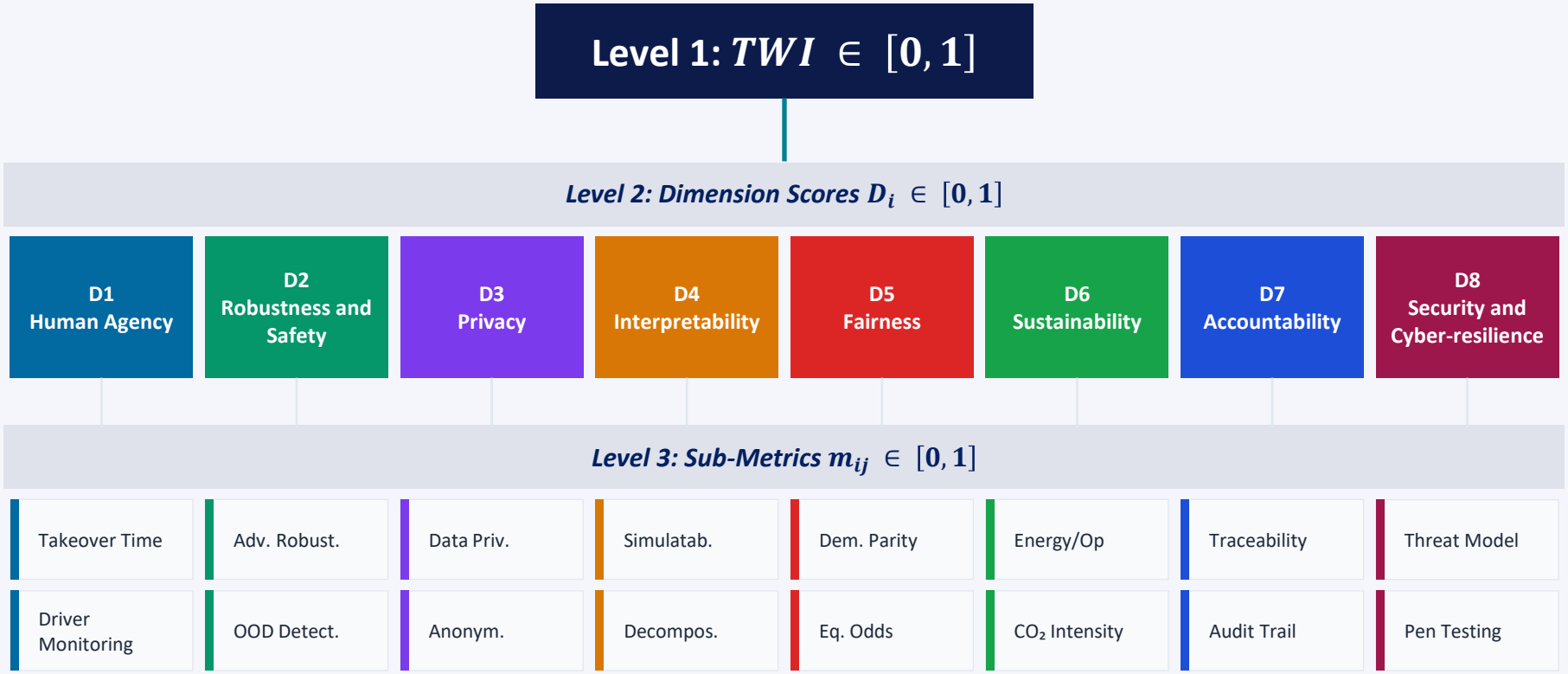
# GAP ANALYSIS

## Standards vs. AI Trustworthiness Requirements

| Dimension | ISO 26262 (Automotive) | IEC 61508 (Industrial) | IEC 62304 (Medical) | DO-178C (Aviation) | EU AI Act |
|---|---|---|---|---|---|
| D1: Human Agency | M | M | L | M | M |
| D2: Robustness and Safety | C | C | C | C | C |
| D3: Privacy | L | L | M | L | M |
| D4: Interpretability | M | M | M | C | M |
| **D5: Fairness ⚠** | N | N | M | N | L |
| **D6: Sustainability ⚠** | N | L | N | L | L |
| D7: Accountability | C | C | C | C | C |
| D8: Security and Cyber-resilience | C | C | M | C | C |

Coverage: C Comprehensive M Moderate L Limited N None

# TWI Architecture: 3-Level Hierarchy



# FORMULATION

## Lifecycle Evidence Integration

### $E_p$ Process Evidence

ISO 26262 traceability, hazard analysis, design documentation

### $E_v$ Verification Evidence

Benchmark metrics (KITTI, nuScenes), test reports, formal V&V

### $E_o$ Operational Evidence

Field incident rates, uptime, runtime monitoring statistics

$$m_{ij} = w_p E_p(D_i) + w_v E_v(D_i) + w_o E_o(D_i)$$

$(w_p + w_v + w_o = 1$, vary by domain maturity)

## TWI Formula

Dimension scores via weighting sum of sub-metrics:

$$D_i = \sum_{j=1}^{k_i} \beta_{ij} m_{ij}$$

Non-compensatory gate:

$$G_{domain}(D) = \begin{cases} 1 & D_j \geq \theta, \forall j \in C_{domain} \\ 0 & otherwise \end{cases}$$

$$TWI = G_{domain}(D) . \sum_{i=1}^{8} w_i D_i$$

*Non-compensatory gate: critical dimensions must ALL pass their threshold θ — failure in any critical dimension sets TWI = 0*

**Automotive**

Critical Dimensions: D1, D2, D8

Min. Threshold: θ = 0.70

**Industrial**

Critical Dimensions: D2, D6, D7

Min. Threshold: θ = 0.75

**Medical**

Critical Dimensions: D2, D4, D5

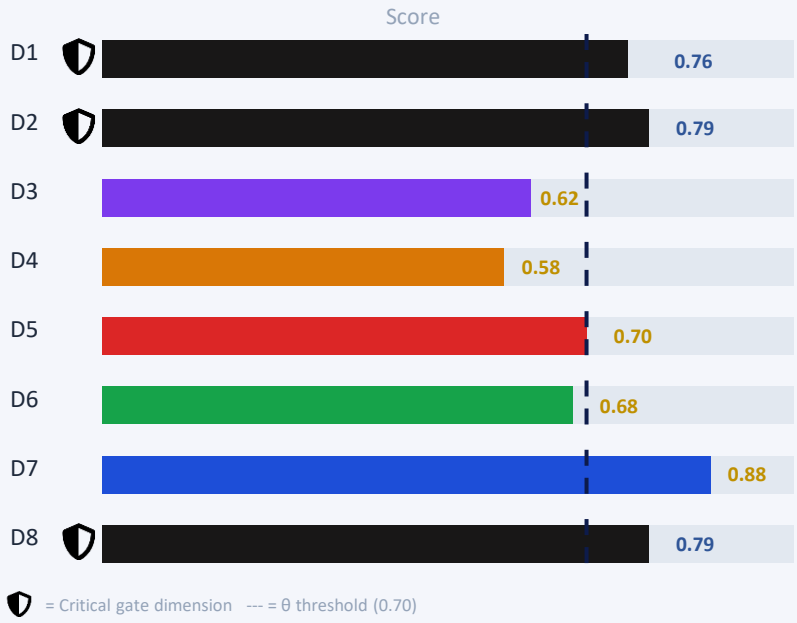
Min. Threshold: θ = 0.80

# APPLICATION

## Automotive Level 3 ADS — Illustrative Assessment

### AHP Weights

| | Dimension | $w_i$ |
|---|---|---|
| D1 | Human Agency | 0.18 |
| D2 | Robustness and Safety | 0.16 |
| D3 | Privacy | 0.15 |
| D4 | Interpretability | 0.14 |
| D5 | Fairness | 0.11 |
| D6 | Sustainability | 0.10 |
| D7 | Accountability | 0.09 |
| D8 | Security and Cyber-resilience | 0.07 |




**TWI = 0.7462**

✓ Acceptable

**Strength: D7 Accountability (0.88)**

ISO 26262 compliance, traceability, mature safety cases

**Critical Gap: D4 Interpretability (0.58)**

Post-hoc saliency maps ≠ real-time interpretability for drivers

**Gate Margin: D1 at 0.76 vs θ=0.75**

Only +0.01 margin at higher threshold; priority for improvement

# SENSITIVITY ANALYSIS

**Baseline TWI**

**0.7462**

Automotive L3

**Monte Carlo (10K)**

**0.7224**

Mean | 95% CI: [0.686, 0.759]

**Weight Sensitivity**

**±0.14%**

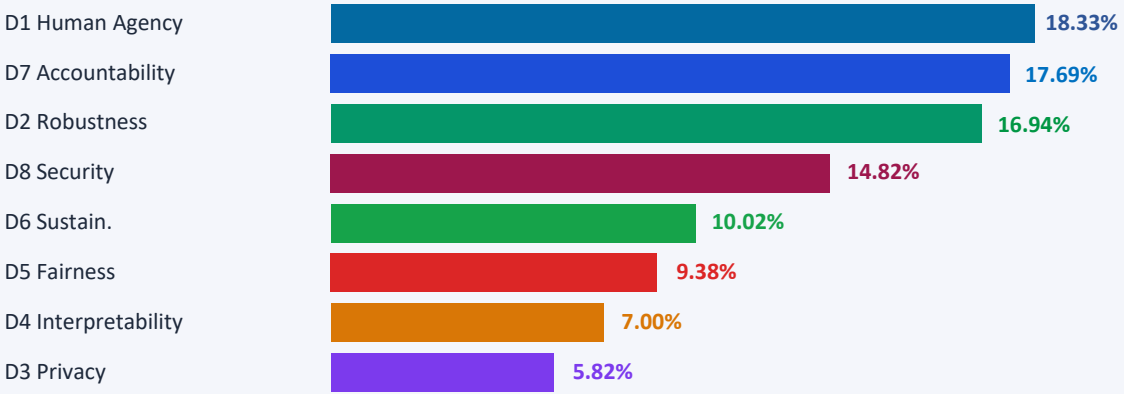
Proportional redistribution

**Extreme Perturbation**

**±3.62%**

$w_1$: 0.10 → 0.25 shift

## Dimensional Contribution to TWI




**Top 3 Dimensions**

account for

**53.0%**

of total TWI

# Priority Shift Across Application Domains

## Automotive L3 ADS

Top Priority:

**Human Agency**

| D1 | D4 |
|---|---|
| 0.76 | 0.58 |
| Human Agency | Interpretability |

**D4 Interpretability = 0.58 — biggest gap**

Post-hoc XAI insufficient for real-time driver decisions

## Industrial — Steel Defect

Top Priority:

**Sustainability**

| D2 | D6 |
|---|---|
| 0.975 | 0.50 |
| Robustness | Sustainability |

**D2 Robustness = 0.975 — excellent**

24/7 operation: ~12 kWh/day/station → sustainability critical

## Medical — CXR Diagnosis

Top Priority:

**Fairness**

| D2 | D5 |
|---|---|
| 0.84 | 0.65 |
| Robustness | Fairness |

**D5 Fairness = 0.65 — critical gap**

High AUC (0.84) coexists with systematic underdiagnosis bias

**Key insight: same 8 dimensions, different priorities**

# CONCLUSIONS

## What We Contribute

- Multi-dimensional TWI framework bridging domain safety standards (ISO 26262, IEC 61508, IEC 62304) and horizontal AI regulations (EU AI Act)
- Lifecycle evidence integration: process + verification + operational
- Non-compensatory critical gates prevent safety failures from being masked by high scores elsewhere
- Domain-adaptive without structural change; used same 8 dimensions, shifted priorities

# LIMITATIONS & FUTURE WORK

## Limitations

- Illustrative validation only
- Expert estimates and synthetic scenarios
- Uncalibrated domain thresholds
- Incomplete theoretical foundation

## Future Work

- Real-world validation
- Stakeholder-based AHP
- Threshold calibration
- Formal convergence proofs and alternative aggregation

# Research Progress

- Toward a Trustworthiness Index for AI-Enabled Safety-Critical CPS: A Multi-Dimensional Framework
  Accepted and presented in 2026 European Conference on Reliability, Availability, Maintainability, and Safety (RAMS-Europe 2026)

- A Composite Human Agency–Fairness Index for Trustworthy AI in Cyber-Physical Systems
  Submitted to IEEE Transactions on Artificial Intelligence

- Robustness and Interpretability Indices for Trustworthy AI
  In progress

NTNU | Norwegian University of Science and Technology

# Thank You

**Contact: dina.t.utari@ntnu.no**

Norwegian University of Science and Technology (NTNU), Trondheim, Norway