CONNECT

CCAM TRUST & RESILIENCE

Challenges and Priorities towards Trustworthy AI

WORKSHOP REPORT

License	Creative Commons 4.0 International - With Attribution
Date	4 September 2025
DOI	10.5281/zenodo.16893415
Editors	Ioannis Krontiris, Huawei Technologies, Germany Frank Kargl, Ulm University, Germany Thanassis Giannetsos, Ubitech Ltd., Greece

https://horizon-connect.eu

This document was typeset with LaTeX.

Funded by the European Union under Grant Agreement No. 101069688. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or CINEA. Neither the European Union nor the granting authority can be held responsible for them.

The information in this document is provided as is, and no guarantee or warranty is given that the information is fit for any particular purpose. The content of this document reflects only the author's view – the European Commission is not responsible for any use that may be made of the information it contains. The users use the information at their sole risk and liability.

Foreword

The rapid advancement and widespread deployment of artificial intelligence systems across critical sectors has created an urgent imperative for ensuring their trustworthiness, reliability, and alignment with fundamental rights and societal values. As AI technologies become increasingly embedded in domains such as 6G networks, Connected, Cooperative, and Automated Mobility (CCAM), healthcare, and critical infrastructure, the importance for AI trustworthiness has never been bigger. Against this backdrop, the European Union's regulatory landscape is evolving rapidly, with frameworks such as the EU AI Act [1] establishing new requirements for AI system governance, transparency, and accountability.

Recognizing these pressing challenges, the CONNECT project [2] organized a comprehensive Workshop on Trustworthy AI on March 25-26, 2025, in Frankfurt, Germany. Co-organized with the REWIRE project [3] and involving multiple European AI research initiatives, the workshop brought together a group of leading experts from across Europe, including AI researchers, policymakers, industry stakeholders, and standardization experts, in order to collaboratively surface gaps in current methods, share lessons from use cases and different domains, as well as debate future priorities.

Workshop Motivation and Objectives

The workshop was designed around the understanding that achieving trustworthy AI requires a holistic approach that addresses data governance, regulatory compliance and ethical alignment throughout the entire AI system lifecycle.

More specifically, on the first day the workshop was structured around three focused panel sessions:

- **Panel 1: Trustworthiness Assessment of Data in CCAM**: Addressing fundamental challenges in data completeness, representativeness, accuracy, and bias mitigation that directly impact the reliability and fairness of AI systems. The workshop examined how data quality serves as the epistemic backbone of trustworthy AI, with particular attention to the unique challenges posed by distributed, multi-stakeholder environments.
  Panelists:
  - Michael Buchholz — Ulm University, *PoDIUM* project
  - Karla Quintero — IRT SystemX, *AI4CCAM* project
  - Lakshya Pandit — Rupprecht Consult, *AITHENA* project
  - Fouad Hadj Selem — VEDECOM, *SUNRISE* project
  - Antonio Kung — Trialog, *CONNECT* project
- **Panel 2: Embedding Trustworthiness Across AI System Lifecycle**: Exploring systematic approaches to integrating trustworthiness properties throughout the complete AI development and deployment lifecycle.

Panelists:
  - Prof. Dr. Birte Glimm — Institute of Artificial Intelligence, University of Ulm
  - Prof. Roberto Navigli — Sapienza University of Rome
  - Dr. Ansgar Koene — Global AI Ethics and Regulatory Leader, Ernst & Young
- **Panel 3: AI Trustworthiness in 5G/6G**: Examining real-world deployment contexts in next-generation networks and intelligent transportation systems, where AI trustworthiness intersects with critical requirements for safety, reliability, low latency, and real-time decision-making under uncertainty.
  Panelists:
  - Mattin Elorza — *6G-OPENSEC* project
  - Prof. Dr. Ghassan Karame — Ruhr University Bochum, *REWIRE* project
  - Mohammed Alfaqawi — VEDECOM, *SUNRISE* project
  - Dr. Thanassis Giannetsos — Ubitech, *CONNECT* project

The second day of the workshop shifted from analysis to co-creation. Building on the insights surfaced during the first day's panel discussions, participants engaged in structured roundtable sessions designed to identify actionable research challenges and shape the foundation of a forward-looking roadmap. Each roundtable was centered around one concrete and provocative question corresponding to the following ones:

- What are the most important open challenges for trustworthy AI and derived research requirements?
- What incentives can be created to make trustworthiness a feature companies actively desire to build and promote, not just because they are required to, but because it strengthens their market position, reduces risk, and increases value?
- How can AI developers and stakeholders anticipate and adapt to evolving regulatory frameworks while contributing meaningfully to their design and operational alignment?
- How can AI resilience be systematically defined, measured, and maintained across lifecycles and operational contexts, so that trustworthiness can be preserved even under uncertainty, degradation, or adversarial conditions?

The output is a concrete map of the most pressing challenges and knowledge gaps in building trustworthy AI systems. By anchoring the discussion in latest engineering practices and sector-specific constraints, this report offers a roadmap that can inform both research agendas and policy development. It serves as a reference point for aligning future European R&D efforts with the societal and operational demands of deploying AI responsibly, especially in critical domains such as mobility, communications, and public infrastructure.

Report Structure and Outcome

This report documents the comprehensive outcomes of the workshop discussions, structured to reflect both the formal presentations and the collaborative working sessions. Part I presents the insights from the three panel discussions, covering trustworthiness assessment in CCAM, lifecycle embedding of trustworthiness, and AI trustworthiness in 6G networks. Part II documents the outcomes of the collaborative round-table discussions, addressing research requirements, incentives for trustworthy AI, regulatory landscape understanding, and AI resilience challenges.

Each section identifies critical research gaps and open challenges that require continued attention from the research community. The workshop's primary outcome is to bring this to a comprehensive report for AI trustworthiness that serves as both a strategic assessment of the current landscape and a call to action for coordinated European research efforts.

Contents

Contents

Executive Summary

AI in Connected, Cooperative, and Automated Mobility (CCAM) sets the hardest benchmark for trustworthiness: safety-critical decisions among many interacting agents, partial and sometimes conflicting evidence, and millisecond control budgets across the vehicle–infrastructure–edge continuum. In this setting, models negotiate with other agents, inherit uncertainty from upstream perception, and operate under non-stationary conditions where the Operational Design Domain (ODD) is continuously tested at the boundary. AI Trustworthiness therefore has to be a system-level invariant maintained through measurable evidence, compositional reasoning across trust domains, and runtime assurance that adapts as conditions change.

Within this dynamic context, the CONNECT workshop on Trustworthy AI took place in Frankfurt on 25–26 March 2025, convening over twenty leading experts from research, industry, and standardization. The two-day event was deliberately structured: the first day unfolded through focused panels on data trustworthiness, lifecycle integration, and AI in 6G, while the second day centered on collaborative roundtables that probed research gaps, incentives, regulatory alignment, and resilience under uncertainty. The discussions collectively mapped a set of engineering imperatives, fundamental research challenges, and standardization priorities that not only chart a path for trustworthy AI in safety-critical CCAM systems, but also offer transferable lessons for the wider deployment of AI in critical infrastructures.

Engineering Imperatives

ODD definition must be evidence-driven ODD definition must be evidence-driven, not declarative. The workshop emphasized that ODDs should be substantiated by scenario-based evidence covering routine, boundary, and rare cases, and rendered operational by encoding quantitative thresholds and trust signals that signal when the system is within, approaching, or outside its boundaries. This enables adaptive control or safe degradation. ODDs should also be treated as evolving constructs, continuously updated through incident reports and evidence of model drift, and maintained in harmonized formats to support sharing and alignment across CCAM stakeholders.

Runtime Trust as a Dynamic Orchestration Parameter A second engineering insight is that trustworthiness in CCAM and 6G cannot be confined to design-time certification but must be enforced at runtime. Trust must operate as an orchestration parameter: explicit service-level agreements that encode required guarantees, continuous evidence gathering through attestation and behavioral signals, and control loops that admit, steer, or shed AI workloads depending on current trust levels. This reframes assurance from a static precondition into a live contract that is monitored, negotiated, and acted upon across domains and over time, which is essential when safety and latency constraints converge at the vehicle–edge–cloud continuum.

Uncertainty quantification as an engineering primitive The workshop recast robustness as the capacity to measure, propagate, and act on uncertainty across the pipeline. Uncertainty must be quantified at input, at inference (well-calibrated confidences, selective prediction/abstention), and at decision (risk budgets and thresholds that trigger policy switches, handover, or safe degradation). These signals must not vanish at system boundaries: they should be encoded as metadata, propagated across vehicle–edge–cloud interfaces, and logged for lifecycle management. CONNECT's use of subjective logic illustrates how such quantification can fuse heterogeneous evidence and produce interpretable trust scores, turning uncertainty into a control primitive rather than a side effect.

Cross-layer trust in 6G architectures The workshop highlighted that 6G will not host monolithic AI systems but chains of models deployed across heterogeneous trust domains. Trustworthiness therefore depends on composability: models must expose interfaces for uncertainty and provenance so that downstream components can reason about trustworthiness. In distributed ML, particularly federated and split learning, the challenge is not just efficiency but ensuring that local training data of varying quality does not erode global trust. The group emphasized that trust signals must be embedded into orchestration itself, influencing resource scheduling, placement, and fallback, rather than being post-hoc metrics. This requires new cross-layer mechanisms where uncertainty and trust metadata are treated as core element in the 6G control plane, enabling AI-driven services to degrade gracefully, interoperate across trust domains, and remain accountable even as components evolve independently.

Resilience through explicit contracts, KPIs, and monitoring infrastructures The workshop emphasized that resilience in AI cannot be reduced to model performance in isolation but must be framed at the system level, with AI embedded in complex socio-technical settings. Trustworthiness therefore requires a semantic contract between lifecycle actors, making explicit the system's boundaries, assumptions, and expected behaviors under uncertainty. This must be backed by multidimensional KPIs that integrate technical robustness (adversarial resistance, graceful degradation, recovery times), operational performance (availability, responsiveness), business impact, and trust metrics. Finally, resilience depends on continuous monitoring infrastructures that combine performance tracking, compliance verification, anomaly and drift detection, and bias monitoring, enabling systems to detect and respond to failures before they propagate to end users.

Research Challenges

Advance neuro-symbolic reasoning Current language models generate outputs from correlations rather than structured reasoning, leaving them prone to fluent but incorrect results in high-stakes settings. Research must integrate explicit, verifiable knowledge, i.e. ontologies, logical rules and symbolic KR, into AI pipelines, so outputs become accountable, auditable, and grounded in reasoning rather than pattern-matching.

Credible signaling mechanisms for users A critical barrier to adoption is the gap between complex system properties and user understanding. Research must develop interpretable trust signals—certification labels, badges, or trust scores that transparently convey fairness, robustness, privacy, and accuracy in ways non-experts can grasp. Such mechanisms must balance simplicity with completeness, avoid false reassurance, and be anchored in credible certification processes rather than self-assessment, so that they genuinely support informed trust decisions and responsible adoption.

A compositional "evidence calculus" for trust The report repeatedly exposes a missing foundation: we lack a mathematically grounded way to compose trust claims across pipelines, components, and trust domains while carrying uncertainty. Research should define a semantics where assumptions, guarantees, calibration quality, data lineage, and operating context can be combined and updated with sound rules. Think of it as an "evidence calculus" that supports assume-guarantee reasoning under uncertainty, proves when system-level trust holds, and quantifies how individual weaknesses degrade end-to-end assurance. Without this, multi-agent systems can neither argue nor maintain trust coherently over time.

Trust-aware arbitration for multi-agent conflict In CCAM, multiple vehicles may face the same emergency yet produce conflicting AI-driven responses, exposing the limits of assuming uniform models or inputs. The workshop stressed that trustworthiness in such settings depends less on enforcing model homogeneity and more on enabling agents to authenticate, signal their operational boundaries, and exchange structured confidence and reliability profiles under tight latency and bandwidth constraints. This reframes arbitration as a research challenge in trust-aware multi-agent negotiation, where correctness cannot be fact-checked in isolation but requires protocols for reconciling divergent outputs amid uncertainty, ensuring both accountability and safety.

Priorities for Standardization & Regulatory Frameworks

Certification must evolve toward purpose-driven and dynamic assurance Traditional static certification is inadequate for AI systems that adapt, interact, and evolve. The workshop emphasized that certification should no longer be tied only to architecture or components, but to the declared purpose of the system and the conditions under which it operates. This requires purpose-specific assurance test cases that integrate technical metrics with societal values such as fairness, robustness, and safety. Certification must also be continuous: systems remain valid only as long as trust metrics like calibrated uncertainty or interpretability fidelity remain within predefined bounds. Achieving this vision demands shared taxonomies of trust, uncertainty, and risk, and frameworks that generate test cases aligned with operational and societal expectations. Without such a shift, certification risks lagging behind the dynamic reality of AI.

Trustworthiness profiles are emerging as the linchpin for operationalizing AI assurance They translate abstract principles into domain- and task-specific requirements, specifying not just what counts as trustworthy behavior, but how it must be evidenced, verified, and communicated along supply chains and to end-users. In sectors such as automotive or energy, profiles ensure that expectations around robustness, explainability, and resilience are consistently understood by OEMs, suppliers, and integrators. Critically, profiles must remain living instruments, updated as systems evolve, and be made communicable: formally to value-chain partners via technical assurance, and simply to end-users through clear signaling. Without standardized formats and validation processes, however, profiles risk fragmentation and opportunistic misuse, undermining both regulatory compliance and public trust.

Incident reporting is not just an accountability mechanism but a systemic learning infrastructure for trustworthy AI Workshop participants stressed that without structured and verifiable reporting—covering failures and near-misses alike—critical domains such as automotive, healthcare, and infrastructure cannot accumulate the evidence needed to improve resilience. The challenge is to move from fragmented, reactive logging to standardized, interoperable frameworks that enable cross-sectoral sharing of real-world anomalies and edge cases. Such frameworks must be harmonized around clear definitions of what counts as an incident, adopt schemas that ensure comparability across systems, and incorporate verification mechanisms to prevent underreporting. To function effectively, they must also be integrated into deployment workflows and incentivized through both regulatory mandates and reputational benefits. In this way, incident reporting becomes more than compliance—it becomes the backbone of continuous adaptation, collective learning, and trust calibration across AI ecosystems.

1 Introduction

While CONNECT primarily focuses on assessing the trustworthiness of actors and data in V2X, the CCAM Strategic Research and Innovation Agenda (SRIA) makes clear that the end goal is trustworthy automated decision-making, where AI systems become fundamental to CCAM deployments [4]. For that goal to materialise, the *uncertainty* of trust sources must be accounted for both at run time and during AI model training and operation. CONNECT's dynamic trust characterisation offers precisely these building blocks: evidence-based trust claims that can be propagated beyond per-message checks to inform how training datasets are curated and how models later adapt and behave. In this sense, CONNECT provides a CCAM-specific foundation that is compatible with broader trustworthy-AI practices now emerging in the CCAM community [5].

In parallel, CCAM use cases are distributing AI across the vehicle–infrastructure–edge continuum for collective perception, cooperative decision-making, and actuation, work that must reconcile trust with latency and resource limits. This architectural shift also aligns with the Software-Defined Vehicle vision, where data and AI services are orchestrated across vehicle, edge, and cloud. In this context, CONNECT's trust signals become a scheduling and safety primitive, guiding decisions on *what* to use, *where* to compute, and *when* to defer, consistent with emerging work on trustworthy edge intelligence [6, 7]. So naturally there is a bridge from trustworthy data to trustworthy AI models as a forward path for CONNECT, in order to articulate the open challenges that must be addressed and the ways in which CONNECT's building blocks can contribute.

This brings us to the concept of robustness. In conventional machine learning, robustness is narrowly defined as the model's stability under small, adversarial perturbations. However, in CONNECT we also approached robustness as a core KPI for evaluating the Trust Assessment Framework (TAF) itself, ensuring that trust inferences remain stable under noisy, incomplete, or even conflicting V2X inputs. Beyond this, an open challenge remains: how to elevate robustness towards helper assertions on the correctness of AI model outputs, particularly in the context of uncertainty. In safety-critical domains like CCAM, this means that the output of the trust assessment mechanism can meaningfully inform and constrain the internal classification or decision-making process of downstream AI components. This was discussed extensively in the AI Trustworthiness Workshop that CONNECT organized in March 2025[1], where it became obvious that achieving robust and trustworthy AI in practice will require such integration, where trust assessments act as epistemic anchors for AI reasoning, enabling AI systems function in the absence of perfect data, reason under uncertainty, and adapt when their sources of evidence shift or degrade.

In this sense, robustness is a reflection of how well the system handles imperfect or contested knowledge. This perspective aligns with CONNECT's approach: rather than assuming clean, well-curated data, it confronts the problem of epistemic uncertainty directly, treating trust as a dynamic variable

[1] https://horizon-connect.eu/workshop-on-trustworthy-ai-2/

inferred from context and source coherence and reliability. By integrating these trust signals before data is consumed by AI components, CONNECT effectively redefines robustness as the capacity to act wisely under uncertainty. Availability of such dynamic trust assessment mechanisms, can avail next-generation vehicle and infrastructure-based environment perception technologies for robust and reliable CCAM operations. This intelligence lies at the heart of sense-think-act systems of CCAM considering the vehicle, the infrastructure, the cloud at-the-edge while guaranteeing security and safety convergence.

This re-framing has the ability to overcome the limitations of traditional adversarial ML techniques as a primary defence strategy. While useful in controlled settings, adversarial training and similar techniques are fundamentally reactive and narrow, tailored to specific threat models. They operate under assumptions, like known perturbation bounds or stable input distributions [8, 9], that rarely hold in decentralized, dynamic systems such as CCAM. More importantly, they ignore a deeper epistemic flaw: AI systems trained on untrustworthy or unverified data cannot be made trustworthy through model-side adjustments alone. Without knowing the trustworthiness of the data, no amount of parameter tuning can ensure reliable decisions.

The AI Trustworthiness Workshop, which CONNECT organized, explored exactly these aspects, in the context of safety-critical systems, like CCAM, but also extended the discussions to the broader field of Trustworthy AI. Taken together, the insights of this report underline a shift towards a systems-level perspective where trust signals, robustness metrics, and orchestration mechanisms form the foundation of dependable CCAM. It is clear that CCAM progress now hinges on the ability to embed large-scale AI into constrained edge hardware, to validate the full action chain from perception through decision-making to actuation, and to balance latency, energy use, and privacy across vehicle–edge–cloud continuums. The workshop outcomes documented here address precisely these bottlenecks. They move beyond generic "AI for CCAM" to specify how trustworthiness can be engineered, measured, and propagated across heterogeneous domains. In this sense, the report provides the methodological and conceptual scaffolding on which the next generation of CCAM projects can be built.

Part I

Panel Discussions (Day 1)

2 Trustworthiness Assessment of Data in CCAM

Panelists:

Michael Buchholz	Ulm University, *PoDIUM* project
Karla Quintero	IRT SystemX, *AI4CCAM* project
Lakshya Pandit	Rupprecht Consult, *AITHENA* project
Fouad Hadj Selem	VEDECOM Institute, SUNRISE project
Antonio Kung	Trialog, *CONNECT* project

2.1 Insights from the Panel Presentations

2.1.1 Safety

Achieving trustworthiness from the perspective of safety requires the ability to rigorously validate safety performance in real-world conditions. For that purpose, the Operational Design Domain (ODD) plays a central role. It defines the set of conditions under which an AI system is expected to operate safely, transforming abstract safety goals into concrete, testable claims. But while the ODD sets the boundary of responsibility, it does not itself guarantee safe operation. We need to move from scope to evidence and be able to validate safety, meaning that we must populate the ODD with realistic and challenging scenarios that simulate how the system interacts with the world, ranging from routine driving tasks to rare edge cases. This is the function of *scenario-based validation*, which has emerged as a cornerstone for assessing AI behavior under diverse and dynamic conditions. Scenario databases, when sufficiently structured and representative, allow performance evaluation to move from intuition to evidence and risk-informed assurances.

Towards this direction, the panelists underscored the pivotal role of scenario-based validation as a means to evaluate system behavior across a diverse range of situations within the ODD. Scenario databases are seen as essential instruments to simulate and assess both common and edge-case interactions, enabling systematic performance evaluation and risk-based safety assurance. Large-scale demonstration and validation actions (especially those supported through federated data-sharing infrastructures) are explicitly prioritized to overcome the current limitations in scenario coverage and standardization. The CCAM Partnership [4] calls for harmonized European methods for scenario definition, interoperability, and impact assessment, recognizing that only through such coordinated efforts can safety validation become a scalable and certifiable process.

The SUNRISE project [10] rethinks safety not as something you test at the end, but as something you monitor and shape continuously through data. Instead of relying on predefined test cases, SUNRISE builds a living library of real-world driving situations by extracting scenarios from raw sensor data like multi-channel occupancy grids. These scenarios are not just used for evaluation, but they also define how the system understands risk. The ODD is split into zones that reflect how well the system is expected to perform: from routine conditions to the edge of failure. Each zone is linked to specific scenarios and updated over time as new evidence emerges. This turns safety assurance into an ongoing dialogue between the system and the world, not a one-time checklist [11]. SYNERGIES [12], building on SUNRISE and HEADSTART, creates the infrastructure for scaling scenario-based safety validation. It is developing a federated European Scenario Dataspace and a marketplace that pools scenarios from across major projects (e.g., L3Pilot, Safety Pool, ADScene) and supports semi-automated scenario extraction using AI.

While scenario databases help us test whether an AI system is safe within its ODD, the AI4CCAM project [13] takes this a step further by using the ODD as a starting point for system design [14]. In this approach, the ODD helps define what the system is for, how it should behave, and what risks it must account for. By linking scenarios, performance goals, and ethical requirements from the outset, AI4CCAM ensures that trustworthiness (particularly safety, fairness, and human oversight) is built into the system. This allows developers to trace how the system's design aligns with its real-world context, making safety and other trust goals more transparent and accountable throughout the lifecycle.

2.1.2 Robustness

While the ODD defines the conditions under which a system is expected to operate safely, robustness determines how confidently the system can perform within those boundaries and how it responds when nearing or crossing them. In practice, no ODD can fully capture the unpredictability of the real world. Edge cases, sensor noise, or novel combinations of familiar inputs are inevitable. A system's trustworthiness therefore hinges not only on staying within its designed scope but also on its ability to remain reliable and self-aware as conditions fluctuate. Robustness reflects the system's capacity to handle imperfect data, ambiguous inputs, or minor faults, and therefore absorb variability without compromising safety.

In this context, *uncertainty quantification* is important in order to identify the grey zones where predictions are made under doubt and supports decisions about fallback behavior or human intervention. Of course, this kind of trust profile must be monitored and updated continuously, especially as systems learn or encounter new scenarios in real-world deployment. In this way, trust becomes an evolving property, shaped by how a system manages the uncertainty that lies between design intent and operational complexity.

The R.U.M. methodology (Robustness, Uncertainty, Monitoring) [15] offers a structured approach for assessing trustworthiness under real-world conditions. It quantifies uncertainty at multiple levels of the AI pipeline, distinguishing between epistemic uncertainty (from model or data gaps) and aleatoric uncertainty (from environmental variability). Rather than relying on averages or rule-based scoring, R.U.M. uses tropical algebra to aggregate indicators in a way that emphasizes the most critical weaknesses in system performance. This enables the construction of an interpretable trust profile that reflects how transparently and reliably the system operates under uncertainty.

2.1.3 Fairness, Transparency, Accountability and Privacy

The AITHENA project [16] is advancing towards trustworthy AI exploring trade-offs among other important properties: fairness, transparency, accountability and privacy. Instead of proposing a fixed set of metrics, AITHENA develops *checklist-based tools* that guide developers, testers, and stakeholders in evaluating these properties throughout the AI lifecycle [17]. These checklists are designed in such a way that they support nuanced assessments and adapt to different use cases, users, and deployment contexts. For example, fairness includes both technical bias mitigation and societal equity in mobility access; transparency leverages explainable AI methods; accountability is linked to traceability and ownership of decisions; and privacy is grounded in GDPR principles. The methodology is tested in realistic CCAM scenarios using a large pool of real and synthetic data, and supports alignment with emerging EU regulations, including the AI Act [1].

2.1.4 Data as the Epistemic Backbone of Trustworthy AI

As we saw in the previous sections, research has advanced the trustworthiness of AI models through properties like safety, robustness, and fairness. However, less attention has been paid to the trustworthiness of the training datasets that underpin those models. Yet empirical studies show that dataset issues like sampling bias, label noise, and privacy vulnerabilities can undermine model fairness, robustness, and interpretability [18]. In critical domains, like CCAM, even subtle flaws in data collection or curation can propagate into harmful downstream outcomes.

While data quality refers to intrinsic properties like accuracy, completeness, and consistency [19], data trustworthiness is not a static attribute of data but a property that emerges from how data is produced, curated, and shared and how much confidence we can place in that process. A dataset may meet quality benchmarks yet remain untrustworthy, for example, if it originates from opaque or untrustworthy sources. The CONNECT project [2] has shown, in multi-agent systems such as connected vehicles, sensor data from different sources may individually appear high quality, but differ in calibration, source reliability, or integrity. To address that, the methodology that CONNECT developed *quantifies the trustworthiness of data* based on actual evidence about various trust properties (e.g., integrity, authenticity, accuracy).

CONNECT's Trust Assessment Framework (TAF) [20], grounded in Subjective Logic, enables systems to reason probabilistically under uncertainty. It incorporates referral-based trust, where opinions are discounted based on the source's credibility, and supports fusion of multiple trust opinions into a consolidated assessment. This allows vehicles to continuously evaluate the trustworthiness of data and other CCAM actors in dynamic, multi-agent environments in order to take safety-critical decisions, but also enables AI systems more broadly to account for uncertainty, source credibility, and quality when using such data for learning.

2.1.5 Standardization Landscape

Trustworthiness in AI also depends on the existence of clear, interoperable, and widely accepted standards. A key insight from the workshop was the recognition that, in increasingly complex AI systems, no single actor can ensure trustworthiness alone. It was emphasized that trust must be co-engineered across the entire ecosystem, including infrastructure providers, platform developers, data owners, service designers, and regulators. To support this, rather than retrofitting conformance at the end of development, systems must be built using common reference architectures, shared vocabularies, and interoperable mechanisms from the outset.

Towards this direction, *systems design patterns* are important, in order to promote reuse, interoperability, and minimal fragmentation. At the center of this effort are architecture patterns and capability profiles, intended to guide consistent implementation of requirements emerging from the EU AI Act, GDPR, Data Act, and other digital regulations. This approach doesn't impose rigidity, but instead fosters modularity, allowing innovation while safeguarding coherence across systems.

2.2 Identified Open Challenges

After the panel presentations, the subsequent exchange with the audience and among experts highlighted unresolved issues that cut across domains and remain critical for advancing trustworthy AI in CCAM. In this subsection we summarize these open challenges, as they were articulated during the debate.

2.2.1 Communicating Uncertainty to End-Users

Uncertainty does not vanish simply because a system operates within its defined ODD. Even when all environmental parameters appear to be within spec, the system may still lack confidence due to hidden biases in the training data, limited exposure to similar scenarios, or internal ambiguity in its decision-making pipeline. The implication is clear: users must be informed not just when the system exits the ODD, but also when it is unsure, even within it. This is where uncertainty quantification becomes more than a technical metric. It becomes a communication challenge. Without mechanisms to convey system confidence in an understandable and actionable way, users may overestimate the system's capabilities, trust it inappropriately, or fail to intervene when needed.

This challenge highlights a critical research gap at the intersection of uncertainty estimation, explainable AI, and human factors. Current methods for uncertainty quantification are largely inward-facing. They are designed for developers and engineers, not for the people who must interact with these systems in real-time. The question of how to communicate uncertainty to users remains unsolved. Simply showing a confidence score or a warning is unlikely to be sufficient. Users need to understand what the system is uncertain about, why that matters, and what actions they might take in response. This is especially important in shared-control systems, like CCAM, where human and machine decisions are interdependent. Ultimately, the ability to communicate uncertainty meaningfully and contextually becomes a precondition for trust.

2.2.2 Monitoring Over Time

A key challenge discussed in the panel was the lack of systematic monitoring of AI systems after deployment. While much attention is paid to design-time testing, validation, and certification, operational-time assurance remains underspecified. Once AI systems are deployed, there is often no structured mechanism to observe how they perform over time, detect deviations, or reassess trustworthiness as conditions evolve. This gap is particularly concerning for systems that update themselves, rely on external data sources, or operate in safety-critical settings. Although emerging standards (e.g. ISO/IEC AWI 42102) recognize the importance of developing methods for both design and operational phases, there is still limited guidance on what to monitor, how often, and who is responsible for ongoing oversight.

A more concrete gap with respect to this is the lack of structured data formats that support traceable, lifecycle-aware evaluation. Without consistent metadata, system logs, or update records, it becomes difficult to reconstruct past behavior, identify regressions, or understand why performance may have degraded. This limits transparency, auditability, and accountability. Related to this, the workshop discussion highlighted the promise (but also the current immaturity) of data cards and MLOps documentation tools. These are often used inconsistently, if at all, and rarely integrated into formal evaluation pipelines. Moreover, as AI systems evolve or learn over time, it remains an open question how to document those changes in a way that preserves trust, aligns with regulatory expectations, and supports future certification or recertification. The absence of monitoring frameworks for dynamic, long-lived AI systems was identified as a foundational gap in building sustained trust.

2.2.3 Structured Scenario Data: The Foundation of Trustworthy AI Remains Unstable

One of the most cross-cutting and persistent challenges discussed in the panel was the lack of a unified, structured approach to scenario data. While scenario-based validation has become a widely accepted methodology for assessing AI safety and performance within a defined ODD, panelists emphasized that there is still no shared definition of what constitutes a scenario, and no standard format for how such scenarios should be represented, shared, or statistically analyzed. Different projects treat scenarios as vectors of numerical values, trajectories, semantic labels, videos, or simulation test cases, each using domain-specific assumptions that hinder interoperability. This fragmentation undermines reuse, benchmarking, and the development of automated tools for scenario generation, coverage estimation, and risk quantification.

A critical limitation in relation to that is the lack of distributional information in scenario datasets. Without understanding the likelihood or representativeness of a given scenario, it is difficult to measure uncertainty or assess robustness in a statistically meaningful way. CCAM systems must develop shared ontologies, metadata schemas, and structured pipelines for real-life scenario collection and continuous updating. This also opens the door to continuous learning, which rely on structured data streams to improve system performance and reduce bias over time. At present, however, the lack of structured, standardized scenario data remains a bottleneck, limiting not only validation and certification, but the very ability to define and monitor trustworthiness across the AI lifecycle.

3 Embedding Trustworthiness Across AI System Lifecycle

Panelists:

Prof. Dr. Birte Glimm	Institute of Artificial Intelligence, University of Ulm
Prof. Roberto Navigli	Sapienza University of Rome
Dr. Ansgar Koene	Global AI Ethics & Regulatory Leader, Ernst & Young

3.1 Insights from the Panel Presentations

3.1.1 Explainability as a Foundation for Trustworthy AI

Explainability is often cited as a pillar of trustworthy AI, yet it remains poorly understood in both research and practice. It is commonly reduced to visualization tools or simplified feature attributions, while its actual function is far broader and more structural. Explainability matters not only for transparency, but also for critical tasks like identifying spurious correlations, exposing misaligned objectives, understanding multi-objective trade-offs, and detecting discrimination. These functions are especially vital in safety-critical systems and societal domains, where misinterpretation can lead to tangible harm. But the standard development pipeline frequently sidelines explainability in favor of model performance, under the mistaken assumption that the two are incompatible or that explanation is an afterthought.

A growing body of research suggests that explanations should be tailored to the needs of different stakeholders, including developers, end users, auditors, and regulators, rather than optimized solely for technical introspection. True explainability must be embedded as a design constraint from the outset, grounded in a clear understanding of how each group is expected to engage with the system. This shifts the focus from simply producing explanations to examining how design choices in architecture, objectives and data influence what kinds of explanations are possible and meaningful.

3.1.2 Designing and Testing for Safety in Multilingual LLMs

Ensuring trustworthiness in large language models (LLMs) increasingly hinges on their ability to behave safely. That is, to avoid generating outputs that are harmful, illegal, or unethical. Achieving this requires full visibility and control over the entire AI development pipeline. The Minerva project [21] addressed this need by training a family of LLMs from scratch on carefully curated, open-source Italian and English datasets. This approach enabled tight control over vocabulary, tokenization, and data composition, which are factors often overlooked when adapting English-centric models. The result was a set of models that not only performed competitively on Italian NLP benchmarks, but also demonstrated greater cultural alignment and linguistic fidelity, which is especially critical for trust in multilingual, real-world deployments.

To evaluate how such models behave under stress and in ethically sensitive contexts, a comprehensive safety auditing framework was introduced: ALERT [22]. This benchmark tests LLMs using over 45,000 red-teaming prompts across 32 finely categorized safety risks, including hate speech, self-harm, criminal advice, and misinformation. Results from testing 10 state-of-the-art models, both open- and closed-source, revealed that even leading systems like GPT-4 and Llama 2 exhibit significant safety vulnerabilities, especially under adversarial prompting. For example, models that appear safe under normal queries often fail under prompt injections or subtle manipulations, exposing a critical blind spot in current safety measures. ALERT's multilingual dimension also uncovered that non-English responses are often less safe, a finding that underscores the importance of language-specific safety evaluation.

3.1.3 Bias Considerations for Trustworthy AI

Bias in AI systems is a structural consequence of every design decision, from data selection to the definition of system objectives. The challenge is not to eliminate bias altogether, but to ensure that any biases embedded in a system are explicitly identified, well understood, and, where appropriate, justifiable. Trustworthy AI requires acknowledging that outcomes will affect different groups in different ways, and that these impacts must be assessed and addressed proactively. A model that performs well for one population but poorly for another cannot be considered fair or trustworthy. This calls for a shift in mindset: developers must continuously ask whom the system is intended to serve and who may be inadvertently excluded or disadvantaged.

To support this, the IEEE P7003 standard [23] offers a structured framework for addressing algorithmic bias. It encourages developers to document key decisions related to data selection, system objectives, and fairness criteria in the form of a ”bias profile”: a structured record of what was chosen, why it was chosen, and how it was evaluated. Far from being mere compliance paperwork, this profile promotes internal accountability and enables external scrutiny. It allows stakeholders such as regulators, auditors, and affected users to better understand how the system was designed and whether its behavior can be trusted. For instance, a product recommendation system would be expected to show how it performs across different user groups, rather than optimizing solely for the majority population.

Another big focus is on setting boundaries and communicating clearly. AI systems are often used in ways their creators didn't expect, and users can misinterpret their outputs as being more fair or accurate than they really are. That's why the standard also asks teams to define where their system can be safely used, and how to guide users in interpreting the results. Ultimately, trust doesn't come from technical performance alone. It comes from transparency, inclusivity, and clear communication with the people AI affects.

3.2 Identified Open Challenges

As we turn to the open challenges, it is important to recognize the growing urgency for action. AI systems are transitioning rapidly from research to real-world deployment, while regulatory instruments such as the EU AI Act are still evolving [1]. Key terms like risk management and robustness remain under-defined, leaving both developers and regulators without shared technical criteria for compliance. While regulatory delay may be inevitable, the definition of safeguards, evaluation frameworks, and assurance practices cannot wait. Establishing these foundations is essential not only for legal clarity, but also for fostering responsible innovation and international alignment.

Rather than replicating the race toward ever-larger foundation models, Europe could take a differentiated approach by developing *smaller, more targeted systems* that are explainable, controllable, and tailored to high-stakes domains such as healthcare and mobility. In these settings, data is often scarce, and transparency is a prerequisite for trust. Some recent models have shown that high performance can be achieved with moderate computational resources when combined with curated datasets and semantically informed architectures. These systems may offer a more sustainable and trustworthy path forward, particularly in regulated or safety-critical environments where alignment, oversight, and adaptation are essential.

One thing to emphasize is that trustworthiness in AI is not an absolute endpoint, but a context-specific and evolving goal. In open-world systems like autonomous driving or general-purpose LLMs, perfect trustworthiness may remain unattainable due to intrinsic limitations, such as unknown data origins, non-deterministic behavior, and the inherent ambiguity of language. Trust, in this sense, must be carefully scoped and managed. AI systems should be designed to communicate their boundaries, handle failure gracefully, and be verifiable to the extent possible within their domain. Users are often capable of dealing with uncertainty or imperfect reliability, as long as their expectations are aligned with system behavior. Indeed, trustworthy AI is not about eliminating all uncertainty, but about building systems that are transparent, resilient, and aligned with human expectations.

3.2.1 Managing Conflicting AI Decisions in Multi-Agent Systems

The panel addressed a pressing question raised from the audience: What happens when multiple vehicles, each running a different AI model, face the same emergency, but produce conflicting responses? In the context of CCAM, this scenario is not hypothetical, but it rather reflects a growing reality where models vary in architecture, training data, and interpretation strategies. Panelists agreed that deploying a single standardized model across all systems is neither realistic nor effective. Besides, even identical models can diverge due to input noise or local conditions. Instead, they highlighted the need for further research into how different models can be used together, particularly how they might communicate and negotiate in real-time, under constraints like latency and bandwidth.

Embedding Trustworthiness Across AI System Lifecycle

The discussion also emphasized the importance of being able to evaluate *the trustworthiness of the output of AI Agents*. Especially in terms of correctness this cannot be reduced to simple fact-checking. For example, in language-based or context-sensitive systems, determining whether a model's output is "appropriate" or "correct" involves more than verifying past facts. It requires understanding the complexity and intent of the generated content. This also becomes especially important in CCAM scenarios, where synthetic or ambiguous outputs can have safety-critical consequences. We currently lack robust methods to verify AI outputs in such settings and more research is needed in this direction. At the same time, we have to be cautious against designing systems based on total trust, meaning that AI-models must be able to function amid uncertainty and to do that, uncertainty must be quantifiable.

In parallel, a closely related open challenge is the lack of mechanisms for AI Agents to explicitly communicate their operational boundaries and reliability profiles. This gap becomes particularly critical in distributed, real-time environments. While much attention is placed on making the "right" decision, equally important is the ability of a system to signal how confident it is in that decision, under what conditions it was generated, and whether it remains valid in the current context. This is not only relevant for human users interpreting system behavior, but also for Agent-to-Agent coordination, where conflicting or uncertain outputs must be reconciled without shared models or centralized control. The ability to communicate limitations, such as degraded sensing, low confidence, or context ambiguity, could allow agents to adjust behavior, defer action, or trigger fallback protocols. The discussion underscored that trustworthy AI in multi-agent systems requires not only sound reasoning, but transparent signaling of system boundaries, so that cooperation can occur even amid imperfect information and partial alignment.

3.2.2 Advancing Trustworthy AI Through Neuro-Symbolic Reasoning

A key challenge raised in the panel discussion is the need to move beyond the limits of purely statistical AI models by advancing hybrid, Neuro-Symbolic reasoning approaches [24]. Current language models, while highly capable, operate by generating outputs based on statistical correlations rather than structured reasoning. As several panelists noted, this makes them vulnerable to producing fluent but incorrect or contextually inappropriate results, especially in high-stakes domains where outputs must be verifiable and accountable. The discussion emphasized that LLMs "can fail greatly," and that incorporating structured, verified knowledge resources, such as ontologies or logical rules, could provide critical grounding for their outputs. This perspective reflects a growing research consensus, notably articulated in the recent Dagstuhl Manifesto on Knowledge Representation [25], which argues that the limitations of today's AI systems can only be addressed by reintegrating explicit, symbolic knowledge into the AI pipeline.

3.2.3 Bridging the Gap Between Technical Correctness and Ethical Alignment

One other open question is the relationship and potential tension between technical AI development and ethical or societal considerations. For example, would it be more productive to allow technical and ethical research to proceed at different speeds, given that technical metrics like accuracy are easier to quantify, while issues like fairness, bias, and cultural values are often context-dependent and harder to converge on globally? Today it is possible to imagine a technically correct system that is commercially successful but misaligned with societal values, prompting the question: Can such a system still be called trustworthy?

While panelists acknowledged that separate research tracks are inevitable and even necessary, they emphasized that true trustworthiness requires integration. Technical solutions must be contextualized within clearly defined social values and use cases; otherwise, a system could be technically sound but ethically misaligned or opaque in its consequences. The challenge lies in building interdisciplinary collaboration (despite its difficulty) and in ensuring that AI systems, even if partially unethical or commercially motivated, can still be assessed, audited, and improved from both perspectives.

4 AI Trustworthiness in 6G

Panelists:

Mattin Elorza Forcada	Telefónica Innovación Digital; 6G-OPENSEC project
Prof. Dr. Ghassan Karame	Ruhr University Bochum; REWIRE project
Mohammed Alfaqawi	Vedecom Institute; SUNRISE project
Dr. Thanassis Giannetsos	UBITECH Ltd; CONNECT project

4.1 Insights from the Panel Presentations

4.1.1 On the Composability of Trustworthy 6G Service Provision

The Compute Continuum (CC) refers to a seamless and dynamic integration of computing, networking, and storage resources across a range of environments, from edge devices and embedded systems to cloud data centers. Unlike earlier models such as fog computing, the CC enables services to be decomposed into smaller components that can be flexibly deployed wherever resources are most available or appropriate, often shifting in real time based on changing conditions. While this makes systems more flexible and responsive, it also creates a major challenge: how do we make sure we can trust every part of this distributed system, especially when it's spread across different domains and security environments? One of the things we need to make sure of, is that trust is managed and assessed continuously, when a service starts, but also as it runs and moves. This means checking which parts of the system are involved, whether they meet the expected guarantees (like privacy or security), and making sure that data flows only through trustworthy paths. To do that we need tools like attestation, policies for how services are allowed to work together, and models that track how trust changes across the system. This approach helps ensure that AI-enabled services can safely operate across complex, real-world networks, where not everything can be trusted by default.

One specific challenge that CONNECT project [2] is addressing is the difficulty of establishing trust in MEC environments, particularly in multi-stakeholder 6G deployments. The project specifically develops the use case of a Slow-Moving Traffic Detection (SMTD), where vehicles offload video streams to nearby MEC nodes for traffic analysis based on machine learning. While this offloading enables low-latency, high-efficiency AI-driven decision-making, it also introduces trust challenges: ensuring that both the vehicle and the MEC infrastructure can be verified before offloading occurs. The CONNECT framework addresses this with constructing evidence-based trust claims, which are dynamically evaluated to maintain security in high-mobility, real-time contexts. Ultimately, the CONNECT architecture promotes trust-aware task scheduling, where sensitive workloads are routed only to verified compute nodes, enabling secure and adaptive AI deployment across fragmented, heterogeneous 6G infrastructures.

4.1.2 Trust evaluation on a multi-domain environment

As 6G moves toward increasingly open, disaggregated infrastructures, where services span multiple independent providers, 6G-OPENSEC project [26] proposes a model in which trust is not only monitored, but actively orchestrated. Here, AI plays a dual role: first, by enabling predictive risk assessment, where patterns in system behavior help anticipate and mitigate potential violations of trust or security guarantees; and second, by powering prescriptive decision-making, where the system can autonomously reconfigure services in response to evolving conditions. This intelligence is tightly integrated with intent-based management, where user or operator expectations, such as privacy, resilience, or compliance, are formalized into machine-readable Service Level Agreements (SLAs). These include both security SLAs (SSLAs) and Trust SLAs (TSLAs), which are continuously enforced via closed-loop control and supported by blockchain-based smart contracts that formalize and audit cross-domain commitments.

The trustworthiness of AI-powered services is no longer seen as a static precondition, but as a runtime quality that must be monitored, negotiated, and adapted across the full service lifecycle. 6G-OPENSEC responds to this need by defining a quantifiable trust model, in which different providers and infrastructure domains are assessed based on their ability to meet required guarantees, ranging from infrastructure integrity and data protection to compliance with AI safety or ethical policies. These assessments feed into a Trust Manager that supports dynamic provider selection and enforces fallback mechanisms if trust is violated. In practice, this means that AI-enabled services like CCAM or remote diagnostics are only deployed across those providers whose real-time trust levels meet the required thresholds. By embedding trust as a first-class orchestration parameter, 6G-OPENSEC moves beyond static defense strategies and toward a more resilient, responsive model for managing AI, where trustworthiness is enforced as part of the service logic itself.

4.1.3 On the Robustness of Distributed ML

Building on the previous discussion about dynamic trust assessment in 6G edge-cloud architectures, the REWIRE project [3] adds an important perspective by focusing on the security and trustworthiness of AI systems deployed in zero-trust, distributed environments. A key insight from the project is that AI models (particularly those trained in centralized settings) remain vulnerable to adversarial attacks, including those where attackers do not need access to model internals. To counter this, REWIRE investigates whether robustness can be improved by decentralizing both the training and inference processes of machine learning systems. Their findings, based on extensive experiments, suggest that distributing training across multiple nodes, each using distinct hyperparameters and data subsets, can significantly increase resilience to transfer-based attacks, where adversarial examples crafted for one model are reused against another.

This exploration into distributed ML is especially relevant in edge-enabled, multi-agent AI ecosystems like those envisioned for 6G, where no single party controls all components, and AI agents may operate with different assumptions, hardware, and datasets. REWIRE's methodology aligns well with this setting: each learner autonomously tunes its model based on local data and parameters, and the ensemble then aggregates outputs in ways that preserve diversity, which is essential for robustness. Notably, their results challenge assumptions about the relative importance of various training factors: hyperparameter tuning had a far stronger effect on robustness than architectural diversity, and distributing the inference phase added further defense against adversarial manipulation. These insights are particularly valuable for building trustworthy AI into connected systems-of-systems, where safety and accountability depend on the ability to prevent single points of failure and limit the transferability of attacks across subsystems.

4.1.4 AI-Driven 6G PHY Layer for CCAM

Extending the discussion on trust and AI beyond orchestration and edge computing, recent work on transformer-based neural receivers highlights the critical role of AI at the most foundational level of the 6G stack: the physical layer. In scenarios like CCAM, where vehicles and infrastructure must communicate reliably in real time, conventional signal processing chains may no longer be sufficient. The proposed TransRx receiver replaces these with a fully AI-based model that learns to interpret wireless signals under dynamic, real-world conditions. Built on transformer architectures, it represents a shift from fixed, rule-based pipelines to data-driven, self-adaptive components that can adjust to environmental complexity [27]. This development aligns with the broader trend observed across the workshop: trustworthiness in 6G systems must be addressed end to end, from AI decision-making in the cloud down to signal interpretation at the radio interface. TransRx illustrates how AI can support this vision, not only by improving performance, but by enabling communication systems that are robust to variability, decentralized in operation, and capable of supporting autonomous services without compromising reliability.

4.2 Identified Open Challenges

4.2.1 Balancing Real-Time AI with Security and Trust Requirements

One of the most pressing challenges concerns the tension between real-time AI decision-making and the additional burden of mechanisms to secure these algorithms and enforce AI trustworthiness in 6G systems. In scenarios like Intelligent Transportation Systems (ITS), where AI models are expected to act within milliseconds, the need to encrypt data, validate its integrity, or verify trust levels across domains can introduce delays that risk rendering the system ineffective. At the same time, many of today's AI models remain inherently insecure, lacking robust protection against manipulation or adversarial attacks. Despite their centrality to modern networks, AI models are often deployed without sufficient understanding of their vulnerabilities, and existing mitigation techniques remain immature.

So, this combination of urgency, performance demands, and systemic vulnerability highlights the need for co-designing AI and security architectures from the ground up. Instead of treating security as a layer added on top of functioning systems, the architecture must anticipate trust requirements as part of its core design. This includes lightweight verification methods, context-aware security enforcement, and infrastructure capable of dynamic, cross-domain trust negotiation, all without compromising the responsiveness that real-world AI applications require.

4.2.2 Challenges in Evidence Exchange and Cross-Domain Assurance

In multi-stakeholder 6G environments, AI systems are expected to operate across a fragmented landscape of vendors, domains, and infrastructure providers, many of whom do not, and often cannot, share detailed internal information. As a result, trust must be inferred from abstract "claims" or "trust levels" rather than concrete evidence, leading to a situation where domains must decide whether to collaborate without direct visibility into each other's systems. This highlights the urgent need for new mechanisms that enable privacy-preserving trust assessment, particularly in contexts where data cannot be shared.

This situation becomes even more complex as AI-enabled services are increasingly composed of microservices from multiple vendors, even within a single domain. Current research and standardization efforts, particularly within IETF and related bodies, are grappling with the absence of a harmonized policy language or unified format for encoding evidence and claims related to AI trustworthiness. Ongoing disagreements about schemes, policy enforcement mechanisms, and data modeling approaches reveal a deep gap in interoperability for trust representation. From a research perspective, this points to the need for new approaches to encode, interpret, and validate trust signals beyond isolated systems and across operational boundaries. Moreover, it underscores the urgency of developing evidence exchange frameworks that preserve privacy and abstraction while remaining rich enough to support robust, cross-domain assurance for AI behavior.

4.2.3 Standardization Without Solutions: Anticipating Governance for AI in 6G

There is currently a tension between the growing regulatory demand for AI governance and the lack of mature, agreed-upon technical solutions to guide it. While regulations like the EU AI Act are beginning to require demonstrable trust, robustness, and security, many of the relevant technologies, such as verifiable AI models or cross-domain trust policies, remain underdeveloped or fragmented. It is difficult to standardize trustworthy AI when foundational questions such as "What constitutes secure AI?" or "How do we verify black-box behavior?" are still open. The danger is that premature standardization efforts might lock in assumptions or models that are not yet validated, risking either regulatory overreach or the entrenchment of immature practices.

At the same time, it is important to anticipate governance. There is a growing consensus that standardization efforts must begin now, not with rigid frameworks, but with common terminology, interoperable formats, and shared use cases that can serve as a foundation for future convergence. Current efforts are scattered across bodies like ISO, IETF, ETSI and TCG, each addressing trust, policy, and security from different angles, often without coordination. This fragmentation makes it difficult to align on even basic elements such as how to encode trust claims, exchange evidence, or define policies across network and AI domains. Without convergence, cross-domain AI systems risk becoming unmanageable or unscalable. So there is clearly the need for greater alignment between technical experts and policy makers, including collaborative roadmapping, shared technical reports, and early-stage certification schemes that allow for AI systems to evolve over time without creating deployment dead ends.

Part II

Round-Table Discussions (Day 2)

5 Research Requirements and Open Challenges for Trustworthy AI

Moderator and Editor: Frank Kargl, Ulm University

The second day of the workshop moved from structured presentations to collaborative co-creation. Participants engaged in rotating roundtable sessions, a format designed to ensure that perspectives from different domains and disciplines were systematically exchanged over multiple rounds. Each table was tasked with addressing one of four guiding questions. This section summarizes the outcomes of discussions at Table 1 where we focused on open challenges for trustworthy AI and derived research requirements. Like with the other discussion groups, four groups of five participants each discussed the same topic and we are summarizing the condensed outcomes here.

5.1 Interdisciplinary Approach and Clear Definitions

A first observation made by many participants throughout the discussions centered on the fact that the term "trustworthiness" can be interpreted in a multitude of different ways and that this is a highly interdisciplinary topic ranging from philosophy, to psychology, into many different technical disciplines like security or safety engineering. It also reaches into neighboring concepts like dependability and resilience.

A technical perspective of when an AI-system can be considered trustworthy differs if seen from a human or technical perspective perspective, but both perspectives need to be linked together as human participants of a CCAM system should have a reasonable and realistic understanding of the (technical) trustworthiness of the system. Linking these two domains will require substantial efforts in interdisciplinary research and strengthening this kind of research is a *first challenge*.

Initially, this requires work on cross-sector taxonomies to have clear and compatible definitions so that one domain can relate to the other and mutual understanding can be built. This involves even so simple questions like what constitutes an "AI-system", in particular when this is embedded into larger systems-of-systems.

Another related domain is that of ethics and a *second challenge* is that research is needed on how to align the behavior of AI-systems with the values and ethical principles of their users in different countries and their cultural backgrounds.

Challenge three raised the question if we actually need to have an understand of the trustworthiness of the AI-system itself or whether it would be sufficient to treat this as a black-box and rather trust that the outputs are trustworthy, because this was tested and certified.

5.2 Socio-technical Challenges

Many discussions in our discussions centered around questions how humans use AI-systems and how both interact as a joint socio-technical system.

So our *fourth challenge* is to extend the way we conduct AI research and evaluate the trustworthiness by a "human in the loop" approach and investigate how people interact with AI-systems.

This extends into questions of what mental models people use to conceptualize AI-systems and what trustworthiness requirements are explicitly or implicitly expected by users of those systems, but also by their designers and on a larger scale by society.

If we then consider the concept of explainable AI, which among others aims to make people perceive AI as more transparent and ultimately more trustworthy, we come to *challenge five*.

Besides working on mechanisms for xAI, we also need to investigate how explanations are provided to users in different contexts and how humans will make use and interpret those explanations. For example, complex technical explanations might not serve their purpose and might not be understandable or even create adverse effects of people questioning trustworthiness of AI. So like in earlier challenges, considering the full picture of xAI is definitely an open challenge.

Last in this category, some approaches to trustworthiness like the CONNECT Trust Assessment Framework [20] assume that the system internally assesses the level uncertainty on which it operates and provides its outputs. *Challenge six* is closely related to challenge five and asks the question if and how this level of uncertainty should be communicated to humans (or even to other technical components in the overall system).

5.3 Technical Trustworthiness of AI-Systems

Given the mostly technical audience, technical challenges for trustworthiness of AI-systems were the most frequent discussion topics.

Challenge seven centered on questions how to quantify and measure trustworthiness in the first place. While approaches like the CONNECT TAF [20] aim to make trustworthiness a measurable property of IT-systems, the nature of trust as a unit remains yet to be understood better. This might mean that providing a metric unit for trust might not be achievable, either because of the subjective nature of trust, or by the fact that evidence for trustworthiness is hard to quantify. Related to trustworthiness of AI-systems, one can question if such an absolute measure is actually needed, or if, for example, a relative comparison of trustworthiness of different systems is sufficient.

Securing AI-systems in adversarial settings is our *eighth challenge*. Also motivated by inputs and discussions from day 1, all participants agreed that this requires more attention and understanding. While AI research for many years and decades focused on benign settings, adversarial machine learning has recently gained substantial traction. But given that this is a rather young research field, AI security still needs to catch up to classical IT security related to, for example, formal proofing techniques, methodologies and available mitigations.

The *ninth challenge* is posed by the fact that, in a real world setting, AI-systems are often just subcomponents of much larger systems-of-systems. So even if the trustworthiness of the AI-system itself would be well understood, we would still need to understand the embedding of AI in overall system and the overall trust architecture. Understanding and evaluating the full system will require additional methodologies and technologies. We might even come to the conclusion that an assessment of trustworthiness in such a highly system is not within our reach for a very long time, in which case an external evaluation in specific scenarios and context and the certification of AI-systems becomes even more important.

Last but not least, in our discussions we identified a *tenth challenge*, namely the development of concrete technical solutions to help us strengthen our understanding of trustworthiness on AI-systems. Here, large research efforts are already underway in domains like explainable AI and IT-security. Some directions were concretely named, like approaches to investigate how trustworthiness (or inversely uncertainty and mistrust) can propagate and back-propagate through neural networks or the integration of knowledge-based reasoning capabilities into ML-systems to make the resulting hybrid AI-systems more trustworthy.

5.4 Operational Domains and Certification in those Domains

Even if all the earlier challenges could be solved, we still need to ensure that approaches to achieve required levels of trustworthiness are actually applied in an effective way which is appropriate for the environment where the AI-system is used. This raises up *challenge eleven* on the assessment, evaluation, governance and regulation, and certification of trustworthiness of AI-systems. The more critical the use of such AI-systems becomes, the stricter should be the regime under which the design, development, testing, and deployment of such system falls. This cannot be left to the discretion of companies and developers alone, but clearly requires a certain level of regulation and certification. In the discussion, one aspect that was brought up was whether a purely design-time evaluation of trustworthiness is actually achievable. To the least, this would require very clear definitions of the operational design domain (ODD) and environment in which the AI-system operates. More likely, and this is *challenge twelve* we will also need develop run-time capabilities by which the overall system can monitor the AI-system and self-assess the trustworthiness level on which it operates. Then, gradual strategies can be found and designed to react to reduced levels of trustworthiness in an appropriate way which would maintain design goals like safety.

Research Requirements and Open Challenges for Trustworthy AI

How to test or benchmark the reliability, robustness, resilience of AI-systems was seen as *challenge thirteen* which sees a need to develop suitable methodologies for evaluation and certification . Regulation of this space is only reasonable once such approaches and tools are available. Given the fast paced nature of AI-systems, but also of our understanding of trustworthiness, such regulation must not be static but flexible enough to evolve with our understanding. In the discussions, different approaches to such regulation were discussed (for example, whether it be more oriented towards goals and outcomes or specific measures), but this definitely merits deeper analysis.

6 Incentivizing Trustworthy AI

Moderator and Editor: Ioannis Krontiris, Huawei Technologies

For years, technology providers have expressed their commitment to developing "trustworthy" or "responsible" AI. Yet, in practice, the motivation for companies to go beyond minimal compliance and invest in the additional steps required to achieve true trustworthiness has remained limited. This gap has become especially apparent in competitive environments where rapid feature deployment and market differentiation are prioritized over robustness, fairness, or accountability. Without regulatory mandates, such as those introduced by the EU AI Act [1], the incentive to invest in long-term trust-building measures has often been overshadowed by the race to release cutting-edge functionality.

This raises a critical question for the future of AI governance and innovation:

- What incentives can be created to make trustworthiness a feature companies actively desire to build and promote, not just because they are required to, but because it strengthens their market position, reduces risk, and increases value?
- How can trustworthy AI be reframed not as a constraint, but as a differentiator, which end users, clients, and stakeholders actively prefer and prioritize?

This roundtable highlighted that promoting trustworthy AI is a matter of aligning incentives across the technology ecosystem. Companies need real incentives to invest in trust, whether to meet legal obligations, reduce risk, or gain a competitive edge. At the same time, users and clients must be able to recognize and value trustworthiness in the systems they choose. For trust to become a practical advantage, it needs to be visible, verifiable, and meaningful to everyone involved. The next section explores the tools that can help make this happen.

6.1 Incentives for Companies to Invest in Trustworthy AI

Despite widespread adoption of artificial intelligence across sectors, investment in trustworthiness remains uneven and often reactive. The development of trustworthy AI-systems that are robust, transparent, fair, and aligned with ethical and legal expectations, requires going beyond functional performance to address deeper social and governance considerations. However, without clear incentives, these efforts are frequently deprioritized in favor of faster innovation cycles or feature-driven competition.

Regulatory Drivers The introduction of regulatory frameworks such as the EU AI Act marks a decisive shift in the landscape. These regulations transform trustworthiness from a voluntary commitment to a compliance requirement. By clearly describing risk categories, mandating documentation and transparency, and introducing penalties for non-compliance, such policies compel organizations to internalize the cost of irresponsible AI deployment.

Market Differentiation Trustworthiness can also function as a strategic differentiator. As AI becomes commoditized, the ability to demonstrate system reliability, ethical alignment, and user safety offers a compelling value proposition. Certification schemes or labeling mechanisms, analogous to food safety labels or energy efficiency scores, can help surface trust-related attributes that are otherwise opaque to end users. Establishing such signaling mechanisms encourages market competition on innovation, as well as on integrity and social responsibility of AI systems. This is particularly relevant in sectors where public trust is fragile and reputational risk is high.

Risk Mitigation and Financial Incentives Investing in trustworthy AI contributes directly to organizational risk management. Systems designed with safety, fairness, and accountability in mind are less likely to generate public controversy, regulatory action, or costly incidents. Trustworthiness reduces operational risk and enhances resilience, particularly in high-stakes applications such as autonomous systems, critical infrastructure, and decision-support in healthcare.

Supply Chain and Ecosystem Pressures As AI is increasingly embedded across complex supply chains, requirements for trustworthiness are beginning to cascade through the ecosystem. Large enterprises and public sector entities are beginning to demand assurances about the trustworthiness of third-party AI components, including vendor certifications, model documentation, and ongoing monitoring capabilities. In this context, trustworthiness becomes a prerequisite for participation in certain markets or sectors.

Governance and Capital Investment Trustworthiness is also emerging as a key governance and investment signal. Companies that embed trustworthiness criteria, such as fairness audits, explainability processes, and oversight structures, are better positioned to attract long-term capital. For investors and acquirers, these features indicate maturity, foresight, and reduced exposure to reputational or regulatory liabilities. As responsible innovation becomes a focal point in ESG (Environmental, Social, Governance) strategies, companies that operationalize trustworthiness can strengthen their strategic alignment with both public policy and shareholder expectations.

6.2 Incentives for End Users

For trustworthiness in AI to have meaningful impact, it must be recognized also by those who ultimately interact with AI systems or are affected by their outcomes. However, in many contexts, end users lack the tools, information, or motivation to actively seek out AI systems that are designed with trustworthiness in mind.

Incentivizing Trustworthy AI

Assurance Through Oversight In domains where the risks of failure or harm are high, such as healthcare, financial services, and mobility, users are more likely to turn toward systems they perceive as trustworthy. In these settings, trust can become a deciding factor in whether a user adopts or continues using an AI-enabled product or service. However, the need for trust does not always translate into a need for technical explanations. Rather than understanding every decision an AI makes, many users simply want confidence that robust governance, validation, and monitoring mechanisms are in place to ensure safety, fairness, and compliance. Trust, in this sense, is often contextualized. It may stem from the knowledge that the system has been validated by a regulator, certified by an independent organization, or reviewed by a trusted third party such as a consumer protection group or NGO.

Reliable Disclosure as a Prerequisite for Informed Trust Users and oversight bodies currently lack a reliable way to assess how AI systems perform in real-world conditions or how providers respond to failures. While brand reputation often substitutes for direct evaluation, that reputation depends heavily on what is made visible. Presently, AI incident reporting is fragmented and incomplete. For example, the OECD approach to AI incident reporting [28] primarily relies on media scraping and publicly available sources, resulting in a dataset that is passive, reactive, and coverage-limited. It lacks standardization and depends on whether incidents are reported in the news, leading to both duplication and underrepresentation of less visible but significant failures. In high-risk domains such as autonomous driving, such gaps are especially concerning. Without consistent, mandatory reporting, trust claims cannot be meaningfully verified. This lack of structured visibility undermines public accountability and prevents users, regulators, and institutions from distinguishing between providers with strong safety practices and those without.

The Need for Interpretable Signals to Guide Trust-Based Choices A critical barrier to the adoption of trustworthy AI lies in the mismatch between system complexity and user understanding. While developers and auditors may verify fairness, robustness, or privacy protections, these properties remain largely invisible to non-expert users. This creates a communication gap, where users must decide whether to trust a system without insight into its internal functioning. Most end users lack the expertise to interpret documentation or technical audits, yet still need to make meaningful choices about AI adoption. This information asymmetry undermines both informed consent and appropriate trust decisions. Without clear, credible signals about a system's trustworthiness, users default to indirect signals such as brand reputation, which may not reflect underlying quality or risk. Bridging this gap requires mechanisms that make trust features visible and interpretable to diverse audiences.

6.3 Implementation Mechanisms

To translate the incentives for trustworthy AI into practice, some of the open issues and shortcomings identified in the previous section can be addressed through concrete implementation mechanisms. The following mechanisms represent practical steps discussed in the roundtable for embedding trustworthiness into AI ecosystems in a way that creates alignment, reduces ambiguity, and supports informed adoption.

Trustworthiness Profiles Trustworthiness profiles can be a very useful tool in managing and communicating the multifaceted nature of AI system trustworthiness. These profiles serve a dual function: they establish what constitutes trustworthiness within a specific application context (e.g. autonomous vehicles, medical diagnostics, industrial control systems), and they define how it should be demonstrated, verified, and communicated to others in the value chain or to end users. So in that way, they can serve multiple functions: demonstrate compliance with regulation, articulate added value, and coordinate expectations across supply chains. In sectors like automotive or energy, where AI is embedded into legacy systems, profiles help ensure that trust-related expectations are understood not only by OEMs, but by suppliers and integrators. However, profiles must be task- and domain-specific, and evolve with experience. To function as an effective incentive, they must also be communicable: to value chain partners through technical assurance, and to end users through simplified messaging. Standardization is essential in order to achieve clarity. Without a common language and validation process, profiles risk becoming fragmented or misused.

Incident Reporting Systems The creation of mandatory AI incident reporting systems, particularly for high-risk sectors such as automotive, healthcare, or critical infrastructure, is a critical enabler of trustworthy AI adoption. Such systems would serve as regulatory-grade infrastructures requiring operators to log both failures and near-misses, regardless of whether they are publicly reported. A leading example of this kind of framework is offered by CSET [29], which proposes a structured schema for reporting incidents, including system type, harm category, severity, and mitigation steps. Unlike media-based tracking, this model supports consistent, comparable, and actionable data collection. Ideally, the system would be partially automated and integrated into AI deployment workflows to minimize reporting burdens. For institutional users, it would offer a verifiable source of provider reliability and support risk-based oversight. For companies, participation would signal accountability and build reputational capital, offering a clear incentive to prioritize safety and transparency.

Signaling Mechanisms To make trustworthiness visible and usable for end users, signaling mechanisms such as certification labels, badges or trustworthiness scores are emerging as promising tools. These mechanisms can help address the information asymmetry that typically prevents users from evaluating whether AI systems meet accepted standards of trustworthiness properties, such as fairness, robustness, and accuracy. At the end, these schemes can potentially increase both cognitive and affective trust in AI systems, and in turn, users' willingness to adopt them. However, this effect depends strongly on how the labels are designed. It would be important that labels transparently communicate the criteria on how they were produced and strike a delicate balance between simplicity and completeness. Overly complex labels risk alienating non-expert users, while overly simplified ones may mislead or create a false sense of security. These labels should also be backed by credible certification authorities, rather than relying solely on self-assessment.

7 Understanding and Shaping the Regulatory Landscape for AI

Moderator and Editor: Matthias Schunter, Intel Labs

This section summarizes the discussion in Table 3 that focused on current and future regulations for AI. We collected inputs from 4 groups of 5 participants each. These inputs were categorized and are documented in the subsequent subsections.

7.1 Potential Benefits of Regulating AI

In the early years, no specialized regulation existed to guide AI use. This raised a fundamental question: *why regulate AI at all?* Participants converged on a simple insight: regulation is needed where market incentives and informal good practice are insufficient to prevent harm or to sustain the common resources that trustworthy AI depends on. Two areas stood out.

Managing Risk Regulation can prevent *unacceptable* risks from unmonitored or ungoverned AI deployments. These include unfair or discriminatory decisions and system instabilities, for example when models are trained on AI-generated data and feedback loops amplify errors or bias. A second category concerns decisions that materially affect society but whose quality cannot be verified and whose rationale cannot be explained. Language-based interfaces, particularly when multiple LLMs collaborate, exemplify this: outputs may be non-replicable, opaque, and difficult to audit, which undermines accountability. Well-crafted rules can set preconditions for deployment proportional to impact, such as requirements for oversight, monitoring, reproducibility, and meaningful explanation. Regulation can also reduce legal uncertainty for companies by clarifying permissible uses or indemnifying specific training data categories, thereby enabling innovation while containing liability.

Data Quality Regulation can act on a persistent collective-action problem: high-quality data is costly to produce and maintain, yet its benefits are broadly shared. Without coordination, underinvestment is rational for each actor and overall quality degrades. Targeted measures can mandate cost sharing for critical shared datasets and require minimum standards for provenance, documentation, and refresh rates. This creates durable incentives to contribute to and sustain high-quality resources. A concrete example discussed was real-time traffic maps for mobility management, where shared investment and common quality criteria are essential for safety and effectiveness.

7.2 Potential Risks and Disadvantages of AI Regulations

While regulation can deliver clear benefits, participants cautioned that poorly calibrated rules can create systemic downsides. The deeper insight is that misalignment between regulatory obligations, the technical reality of rapidly evolving AI systems, and market structure can slow learning cycles and skew competition, even when intentions are good.

Reduced Innovation Certification and approval gates can lengthen feedback loops that are essential for progress in AI. If minor updates or patch releases trigger re-certification, teams defer iteration and experimentation, reducing the rate at which models improve and safety issues are discovered. The burden is not only monetary but temporal: elongated review cycles shift research from continuous delivery to infrequent, high-stakes releases.

Increased Cost and Reduced Competition There is a design paradox in regulation. Highly prescriptive rules freeze implementation details and create path dependency, limiting architectural flexibility as methods evolve. Broad, principle-based rules are more future-proof but costly to operationalise, since firms must translate abstract obligations into concrete controls and audits. Large organisations can amortise this "compliance engineering" through scale, while startups, SMEs and research labs face steep fixed costs and expertise barriers. Divergent requirements across jurisdictions further multiply overhead. The net effect can be market concentration and a higher risk of regulatory capture, rather than a diverse, competitive ecosystem.

In short, the risk is not regulation per se, but regulation that is not proportionate, adaptive, and mindful of innovation incentives; getting this balance right is essential to protect society without stalling progress.

7.3 Towards Efficient and Effective Regulations

Participants offered concrete ideas for how future regulations can be both efficient and effective. The common thread was to match obligations to where assurance is most needed, and to design rules that can improve over time.

Scope of a Regulation A central challenge is choosing the right scope. Certification can target *products* (as in common criteria), *processes* (e.g., ISO 9001), or *organisations*. Product certification provides the strongest guarantees for a given system, but it can slow innovation and reduce flexibility. Mandating quality processes helps organisations learn and mature, although early releases may still exhibit limitations. Sector-based regulation is often easier to deploy and to design because specificity allows clearer requirements and enforcement. Including an explicit *expiry or review date* obliges regulators to revisit and recalibrate obligations. In practice, scoping is easier when regulations *start small* by focusing on a specific use case or sector, then expand once evidence accumulates. While the AI Act is an important step, it will need sector-based best practices and profiles that guide each industry in complying with a broad framework.

Continued Learning and Innovation Regulation should allow for experimentation, evaluation, and improvement. Rather than issuing a single "final" text, a draft–refine approach helps align obligations with operational reality. One enabling measure is structured *incident reporting*. As with data breaches or safety events, continuous reporting of AI incidents (for example, vehicle object misclassification) can raise risk awareness and improve deployed systems. Similarly, publishing a *long-term regulatory roadmap* with gradually increasing requirements allows enterprises to plan and to raise guarantees over time, analogous to the Euro emissions series. Clear incentives also matter: positive incentives (such as reduced liabilities for compliant actors) and credible penalties (such as GDPR-like fines for non-compliance) foster uptake.

Access to Training Data Participants noted that it is often unclear which data may be used for which types of AI training and under what conditions. Regulatory requirements also vary by jurisdiction and by the origin of the data, creating fragmentation and legal uncertainty. A recurring recommendation was to harmonize copyright and fair-use rules to simplify lawful procurement of training data across borders. Another idea raised was to ease access for organisations that demonstrably comply with defined safeguards, creating a conditional path to data use analogous to how GDPR facilitates sharing of fully anonymised data.

Transparency The extent to which AI is used inside deployed systems is often opaque to users. Discussions supported increasing visibility of AI use, for example through a clear labelling scheme, comparable in spirit to the Nutri-Score, that distinguishes assistance or correction from fully generated content. An essential part of transparency is *explainability*. For high-impact decisions that affect individuals, participants favoured mandating a level of explanation sufficient for users to understand how a decision was reached and, where appropriate, how it might be contested or improved.

Automotive Regulations In automotive contexts, discussion focused on vehicle homologation and type approval. Current regimes remain centered on mechanical design and safety. As vehicles evolve into "data centres on wheels," an IT-oriented certification approach will be needed. This includes accommodating frequent over-the-air updates without requiring complete re-certification for every change, and addressing the role of external data and data quality in advanced driving functions. Participants also stressed the need for safety and security cultures to converge, so that malicious attacks and probabilistic faults are handled in an integrated manner while maintaining overall safety.

8 Research Requirements and Open Challenges for AI Resilience

Moderator and Editor: Thanassis Giannetsos, Ubitech Ltd.

The increasing integration of artificial intelligence into critical systems across industries has highlighted the urgent need for comprehensive resilience frameworks. Recent high-profile AI failures, e.g. some AI Systems providing dangerous advice or generating false news headlines, demonstrate the gap between current AI capabilities and the robust, reliable systems required for widespread deployment. This section summarizes the outcomes of discussions at table 4 where we focused on open challenges for the essential components needed to achieve AI resilience, addressing the fundamental challenges of standardization, cooperation, and system modeling that currently limit AI trustworthiness.

8.1 Engineering Processes for Trustworthy AI

A foundational insight from the round table was that resilience in AI cannot be meaningfully addressed without rethinking of how AI systems are engineered at the systems level. The traditional focus on improving model performance in isolation must give way to an approach that treats AI as a system being part of a broader, multi-actor sociotechnical setting. So we need processes that go beyond compliance checklists or static verification and they provide a semantic contract between actors across the lifecycle, articulating system boundaries, trust assumptions, and expected behaviors under uncertainty. Currently, these foundational elements—requirements, KPIs, acceptable risk thresholds, and traceable system specifications are often missing.

Establishing clear KPIs for AI resilience requires a multi-dimensional approach that encompasses technical performance, operational efficiency, business impact, and trustworthiness metrics. Technical performance metrics include traditional measures such as accuracy, precision, and recall, but must be supplemented with resilience-specific indicators such as robustness to adversarial attacks, graceful degradation under stress, and recovery time from failures.

Resilient AI systems require comprehensive monitoring infrastructures that provide real-time visibility into system performance and behavior. This monitoring must address multiple dimensions simultaneously: model performance tracking to detect accuracy degradation, operational monitoring to ensure system availability and responsiveness, and compliance monitoring to verify adherence to regulatory requirements.

The monitoring system must be designed to handle the unique characteristics of AI systems, including their probabilistic nature, dependence on data quality, and potential for unexpected emergent behaviors. Advanced monitoring approaches include automated anomaly detection, drift monitoring, and bias detection systems that can identify problems before they impact end users.

8.2 Rethinking Certification: Toward Purpose-Driven and Dynamic Assurance for AI

Traditional certification approaches, designed for static and component-based systems, prove fundamentally inadequate for dynamic AI systems that evolve continuously through learning and adaptation. These frameworks typically assume that a system's behavior is fixed post-deployment and that its components operate independently with clearly bounded functionality. In contrast, AI-enabled systems are dynamic, interdependent, and evolve over time. As a result, a one-time certification snapshot is insufficient. What is needed is a shift toward purpose-driven and dynamic certification models that assess system-wide behavior relative to defined intents and evolving risk thresholds.

At the heart of this shift is a recognition that certification must be tied to purpose, not merely architecture. Rather than certifying individual components in isolation, evaluation should focus on whether the system, as deployed, meets its declared function under operational constraints.

This calls for the definition of purpose-specific assurance test cases, which include not only technical performance metrics, but also contextual values such as fairness, safety, robustness, and compliance. For each purpose or system segment, developers must provide a clear specification of what guarantees are claimed, under what assumptions, and for which classes of users or contexts.

This perspective also demands ongoing validation. As models evolve, the system must be re-evaluated to ensure it continues to satisfy its certified objectives. A system might remain certified so long as trust metrics (e.g., uncertainty estimates, model confidence calibration, or interpretability fidelity) remain within predefined bounds.

Enabling this kind of dynamic certification requires several technical enablers that were identified during the round table discussions:

- Purpose definitions and content models, to structure what is being certified and why.
- Test case generation frameworks for each declared purpose, ensuring that validation reflects both operational conditions and societal expectations.
- Harmonized taxonomies of trust, uncertainty, and risk that allow different stakeholders (developers, users, auditors, regulators) to interpret system behavior consistently.
- Metrics and values as complementary tools: Metrics quantify system behavior, while values define what behavior is acceptable or desirable. Both must be explicitly represented and negotiated across actors.

To support modularity and scalability, the concept of trust profiles was discussed as a foundational building block. A trust profile captures the assurance characteristics of a component or service, i.e., its known behaviors, limitations, trust assumptions, and monitoring hooks. These profiles enable compositional certification, where trust in a larger system can be inferred from the trustworthiness of its constituent parts, assuming their interfaces and interactions are well understood.

8.3 Incident Reporting and Adaptive Monitoring

Participants underscored the necessity of structured, verifiable incident reporting, particularly for AI systems operating in critical domains. Without systematic reporting and feedback loops, the ecosystem cannot learn from failures. Here, the goal is not just reactive logging, but the creation of shared datasets that reflect real-world scenarios, anomalies, and edge cases. These can serve both as test inputs for resilience validation and as training data for adaptive systems.

This insight echoes and builds upon the discussions on the roundtable of Section 6, which emphasized the need for standardized, interoperable indecent reporting frameworks. However, from a resilience perspective, the emphasis shifts toward long-term system learning: reporting mechanisms are not only for accountability but for enabling systemic adaptation. The roundtable called attention to the gap between current reactive logging systems and the proactive, cross-sectoral sharing of real-world failure modes needed for resilient AI.

The roundtable also stressed that incident reporting must be verifiable, incentivized, and integrated into broader governance loops. Without shared protocols for what constitutes an incident, how data should be harmonized, and who owns the feedback loops, lessons from failures remain siloed. This ties directly into the earlier discussion in Section 6 on the importance of creating multi-stakeholder trust frameworks, ensuring that diverse actors can contribute meaningfully to incident learning processes without fear of reputational or legal repercussions.

References

[1] European Commission. *Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206. Official Journal of the European Union. June 2024.

[2] CONNECT Consortium. *CONNECT: Continuous and Efficient Cooperative Trust Management for Resilient CCAM*. https://horizon-connect.eu/.

[3] *REWIRE: Reimagining Trustworthy AI for Future 6G Networks*. https://www.rewire-6g.eu/.

[4] CCAM Partnership. *Strategic Research and Innovation Agenda 2021-2027*. Tech. rep. Jan. 2024. URL: https://www.ccam.eu/wp-content/uploads/2023/11/CCAM-SRIA-Update-2023.pdf.

[5] *Methodology for Trustworthy AI in CCAM*. https://www.connectedautomateddriving.eu/blog/methodology-for-trustworthy-ai-in-ccam/. AI4CCAM blog post. Nov. 2023.

[6] Xiaojie Wang et al. "A Survey on Trustworthy Edge Intelligence: From Security and Reliability To Transparency and Sustainability". In: *arXiv preprint arXiv:2310.17944* (2023). URL: https://arxiv.org/abs/2310.17944.

[7] Pedro Veloso Teixeira et al. *Software Defined Vehicles for Development of Deterministic Services*. 2025. arXiv: 2407.17287. URL: https://arxiv.org/abs/2407.17287.

[8] Yang Song et al. "Constructing Unrestricted Adversarial Examples with Generative Models". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2018. URL: https://papers.nips.cc/paper/8052-constructing-unrestricted-adversarial-examples-with-generative-models.pdf.

[9] Yaniv Ovadia et al. "Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2019. URL: https://proceedings.neurips.cc/paper/9547-can-you-trust-your-models-uncertainty-evaluating-predictive-uncertainty-under-dataset-shift.pdf.

[10] SUNRISE Consortium. *SUNRISE: Safety Assurance Framework for CCAM Systems*. https://www.ccam.eu/projects/sunrise/. 2022.

[11] SUNRISE Consortium. *Requirements for CCAM Safety Assessment Data Framework Content*. Tech. rep. Deliverable D5.1. Horizon Europe Project SUNRISE, 2023. URL: https://cordis.europa.eu/project/id/101103292.

[12] SYNERGIES Consortium. *SYNERGIES: Real and Synthetic Scenarios for CCAM Validation*. https://www.ccam.eu/projects/synergies/. 2024.

[13] AI4CCAM Consortium. *AI4CCAM: Trustworthy AI for Connected, Cooperative & Automated Mobility*. https://www.ai4ccam.eu/.

References

[14] Karla Quintero et al. “Towards a meet-in-the middle approach for Trustworthy AI for CCAM”. In: *8th International Conference on Intelligent Traffic and Transportation (ICITT)*. Firenze, Italy, Sept. 2024. URL: https://hal.science/hal-04758879.

[15] Juliette Mattioli et al. “Leveraging Tropical Algebra to Assess Trustworthy AI”. In: *Proceedings of the AAAI Symposium Series* 4.1 (Nov. 2024), pp. 81–88.

[16] AITHENA Consortium. *AITHENA: Trustworthy AI for Automated Driving*. https://aithena.eu/.

[17] AITHENA Consortium. *Testing and Evaluation Methodology for AI□Driven CCAM Systems*. Tech. rep. Deliverable D5.1. Horizon Europe Project AITHENA, 2024. URL: https://aithena.eu/wp-content/uploads/2024/06/AITHENA-D5.1-Testing-and-evaluation-methodology-for-AI-driven-CCAM-systems.pdf.

[18] Daniel Schwabe et al. “The METRIC-framework for assessing data quality for trustworthy AI in medicine: a systematic review”. In: *NPJ Digital Medicine* 7.1 (2024), p. 203.

[19] ISO/IEC JTC 1/SC 42. *Information technology — Artificial intelligence — Artificial intelligence concepts and terminology*. Geneva, Switzerland: International Organization for Standardization and International Electrotechnical Commission, July 2022. URL: https://www.iso.org/standard/74296.html.

[20] CONNECT Consortium. *Trustworthiness Assessment Framework and Evaluation Methodology*. Tech. rep. Deliverable 3.1. Horizon Europe Project CONNECT, 2024. URL: https://horizon-connect.eu/public-deliverables/.

[21] Riccardo Orlando et al. “Minerva LLMs: The First Family of Large Language Models Trained from Scratch on Italian Data”. In: *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*. Pisa, Italy: CEUR Workshop Proceedings, Dec. 2024, pp. 707–719. URL: https://aclanthology.org/2024.clicit-1.77/.

[22] Simone Tedeschi et al. *ALERT: A Comprehensive Benchmark for Assessing Large Language Models’ Safety through Red Teaming*. 2024. arXiv: 2404.08676 [cs.CL].

[23] *IEEE 7003-2024: Standard for Algorithmic Bias Considerations*. Standard IEEE 7003-2024. IEEE Standards Association, 2024. URL: https://standards.ieee.org/ieee/7003/11357/.

[24] Son Tran, Edjard Mota, and Artur d’Avila Garcez. “Reasoning in Neurosymbolic AI”. In: *arXiv preprint arXiv:2505.20313* (2025).

[25] James P. Delgrande et al. “Current and Future Challenges in Knowledge Representation and Reasoning (Dagstuhl Perspectives Workshop 22282)”. In: *Dagstuhl Manifestos* 10.1 (2024). Ed. by James P. Delgrande et al., pp. 1–61. ISSN: 2193-2433. DOI: 10.4230/DagMan.10.1.1. URL: https://drops.dagstuhl.de/entities/document/10.4230/DagMan.10.1.1.

[26] Centre Tecnològic de Telecomunicacions de Catalunya (CTTC). *6G-OPENSEC_TRUST: DLT-based Trust Management for Open and Disaggregated 6G Networks*. https://www.cttc.cat/project/dlt-based-trust-management-for-open-and-disaggregated-6g-networks/. 2024.

References

[27] Osama Saleem et al. "TransRx-6G-V2X : Transformer Encoder-Based Deep Neural Receiver For Next Generation of Cellular Vehicular Communications". In: *2024 IEEE 100th Vehicular Technology Conference (VTC2024-Fall)*. 2024, pp. 1–7.

[28] Karine Perset and Luis Aranda. *Defining AI Incidents and Related Terms*. Tech. rep. No. 16. Organisation for Economic Co-operation and Development (OECD), 2024. URL: https://www.oecd.org/publications/defining-ai-incidents-and-related-terms-d1a8d965-en.htm.

[29] Ren Bin Lee Dixon and Heather Frase. *AI Incidents: Key Components for a Mandatory Reporting Regime*. Tech. rep. Center for Security and Emerging Technology (CSET), Jan. 2025. URL: https://cset.georgetown.edu/wp-content/uploads/CSET-AI-Incidents.pdf.