



中国汽车工程研究院股份有限公司
China Automotive Engineering Research Institute Co., Ltd.

Agentic AI Trustworthiness Evaluation System

Establishing a Trustworthy Baseline for Intelligent Vehicle Integration

TANG YU, CAERI

2026.09.07

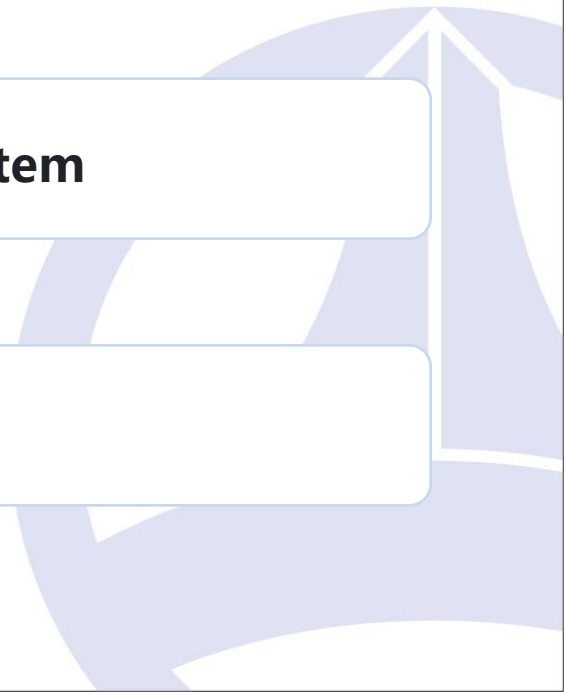
CONTENTS

1 **Agentic AI & Intelligent Vehicles**

2 **EU AI Act Regulatory Framework**

3 **Trustworthy Evaluation System**

4 **Summary & Outlook**



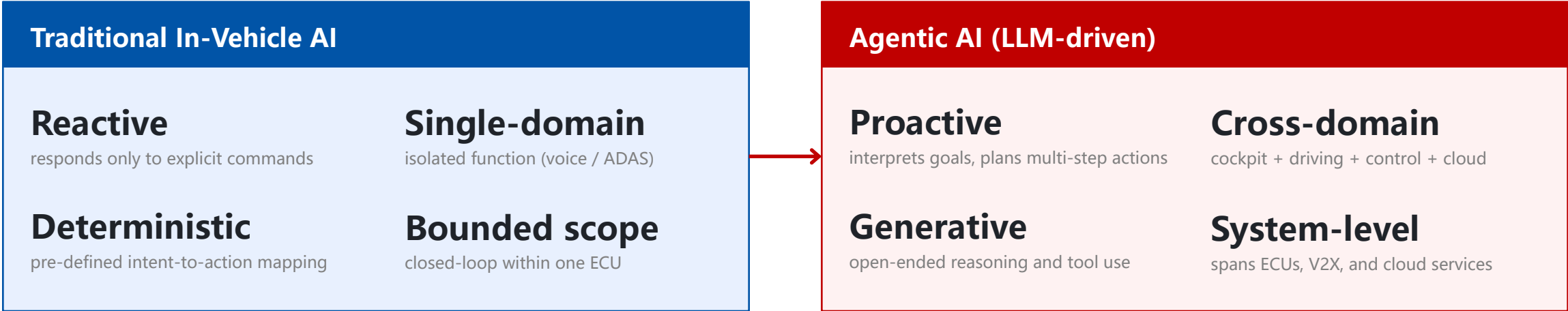
01

Agentic AI & Intelligent Vehicles

To address the risks associated with the in-vehicle Agentic AI in the perception, decision-making, and control chains, this chapter systematically explores how to establish a multi-dimensional trustworthy assessment system.

Agentic AI in Vehicles: Beyond Voice Commands

Agentic AI shifts from passive command execution to proactive goal-driven action across the vehicle stack



CORE AGENTIC CAPABILITIES IN VEHICLES

Perception & Understanding	Reasoning & Planning	Tool Use & Action	Memory & Adaptation
Multimodal scene understanding, driver intent inference, context awareness	Goal decomposition, multi-step planning, trade-off analysis under constraints	Vehicle API calls, device control, service invocation, OTA coordination	User preference learning, contextual memory, continuous adaptation

The shift from reactive to agentic AI expands the attack surface and safety boundary — traditional evaluation is no longer sufficient

The Perception-Decision-Control Loop: Where Risk Emerges

Agentic AI risk propagates across three interconnected loops — each with distinct failure modes

HOW THE LOOP OPERATES

Perception builds the current cabin state.
Decision selects a goal and plan. Control acts through vehicle interfaces.

WHY RISK PROPAGATES

Control outcomes become new perception inputs. Errors can cascade, persist and amplify around the loop.



Perception

- Adversarial sensor input
- Misinterpreted driver intent
- Context hallucination

Decision

- Unsafe goal pursuit
- Flawed planning
- Tool misuse
- Prompt injection

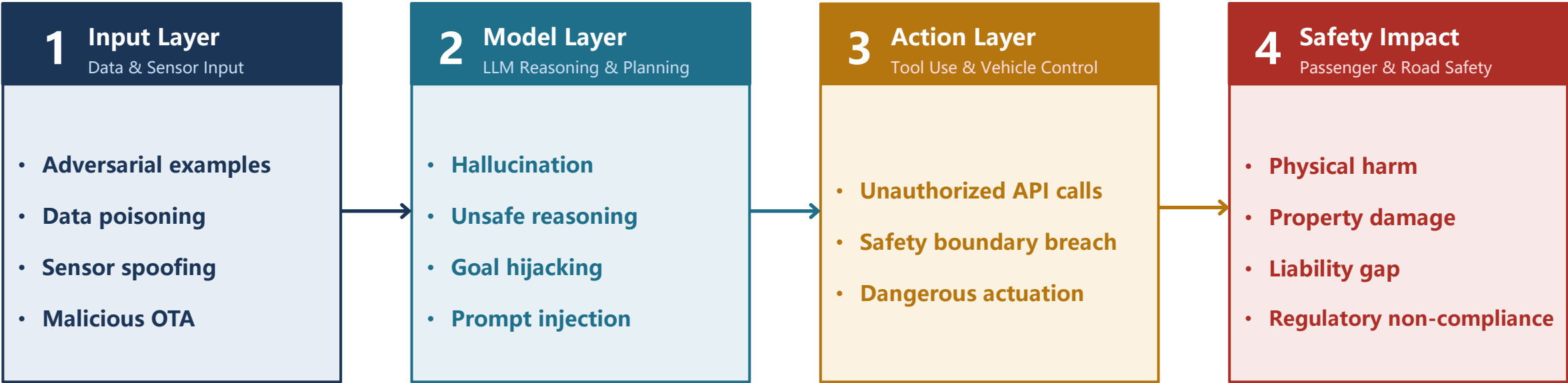
Control

- Unauthorized actuation
- Safety boundary violation
- MRC failure
- Latency

Closed-loop feedback

Four-Stage Risk Chain: From Input to Safety Impact

Risk escalates through four stages — each stage requires dedicated evaluation and mitigation



KEY CHALLENGES FOR EVALUATION



Why Current Evaluation Approaches Are Insufficient

Performance-only evaluation measures what the AI can do — not whether it can be trusted to do it safely

Current: Performance-Based Evaluation

Accuracy and speed

WER, mAP, latency, throughput, feature coverage

Single-domain, static

Isolated function testing, fixed test suites, one-time pass

Benchmark scores

Percentage scores, rankings, competitive comparisons

Blind spots

No safety boundary verification, no adversarial testing
No cross-domain interaction, no liability traceability
No post-market monitoring, no OTA re-certification



Required: Trustworthiness Evaluation

Safety and verifiability

Adversarial & fault injection

System-level, lifecycle

Cross-domain fusion, full evidence chain, continuous assurance

Certifiable grade

Conformity evidence, CAERI + TUV certification

Key advantages

Verifiable safety baselines at integration points
Legally admissible conformity evidence
Actionable pathways for OEM deployment

The industry lacks a universal, compliance-oriented evaluation framework for in-vehicle AI

02

EU AI Act Regulatory Framework

The world's first comprehensive AI law — risk-based classification, high-risk obligations, and enforcement-backed penalties that directly impact automotive AI deployment.

EU AI Act : Four-Tier Risk Classification Framework

Risk determines obligation level — from prohibited practices to voluntary self-assessment



Unacceptable Risk

Prohibited (Article 5)

Social scoring, real-time biometric identification in public spaces, manipulative or exploitative systems, emotion recognition in workplace/education

High Risk

Strict compliance (Annex I & III)

Critical infrastructure, education, employment, law enforcement, migration, justice — and automotive safety components (Annex I). Most in-vehicle AI falls here.

Limited Risk

Transparency obligations (Art. 50)

Chatbots, deepfakes, AI-generated content — must disclose AI interaction and mark generated content. Some cockpit voice assistants may fall here.

Minimal Risk

Voluntary codes of practice

Spam filters, game AI, recommendation systems with no significant impact — no mandatory obligations beyond existing law.

Most automotive AI falls into the "High Risk" category — compliance becomes a mandatory prerequisite for vehicle market access

Annex I safety components + Annex III critical infrastructure = automotive AI under strict regulatory oversight

Automotive AI: Where It Falls in the Risk Classification

Two regulatory pathways bring automotive AI under high-risk obligations — Annex I safety components and Annex III critical infrastructure

Pathway 1: Annex I Safety Components

Scope:

AI systems as safety components of products covered by Union harmonisation legislation, including motor vehicles

Automotive examples:

- ADAS / autonomous driving perception & planning
- Agentic AI controlling vehicle systems (brakes, steering)
- Driver monitoring and safety-critical cockpit functions

Pathway 2: Annex III Critical Infrastructure

Scope:

AI intended to be used as safety component of critical infrastructure, including road traffic management and fleet operations

Automotive examples:

- Fleet management and logistics AI for commercial vehicles
- V2X and connected infrastructure AI systems
- Autonomous fleet operations and robotaxi dispatch

INTERACTION WITH EXISTING AUTOMOTIVE REGULATION

The AI Act complements — not replaces — existing automotive regulations (UN R155 cybersecurity, R156 software update, ISO 26262 functional safety, ISO 21448 SOTIF).

Conclusion: In-vehicle AI predominantly falls under "High Risk" — compliance becomes a mandatory prerequisite for type approval and market access

Annex I (Aug 2028) + Annex III (Dec 2027) = dual regulatory coverage for automotive AI systems

High-Risk Requirements: Mapping to Evaluation Dimensions

Each EU AI Act high-risk obligation maps directly to a testable trustworthiness dimension

EU AI Act Obligation	Article	Evaluation Dimension	Key Test Items
Risk management system	Art. 9	Safety	HARA, dangerous-command ID, MRC verification
Data & data governance	Art. 10	Data Governance + Fairness	Data provenance, bias audit, representativeness
Technical documentation	Art. 11	Transparency	Doc completeness, decision traceability
Record-keeping & logging	Art. 12	Data Governance	Log integrity, retention, audit trail
Transparency & instructions	Art. 13	Transparency	HMI clarity, AI identity disclosure, consent
Human oversight & control	Art. 14	Controllability	Interruption latency, override, E-stop
Accuracy, robustness & cybersecurity	Art. 15	Accuracy + Robustness + Cybersec.	Benchmark metrics, fault injection, pen-test
Transparency to deployers/users	Art. 50	Transparency	Content marking, AI disclosure, watermark

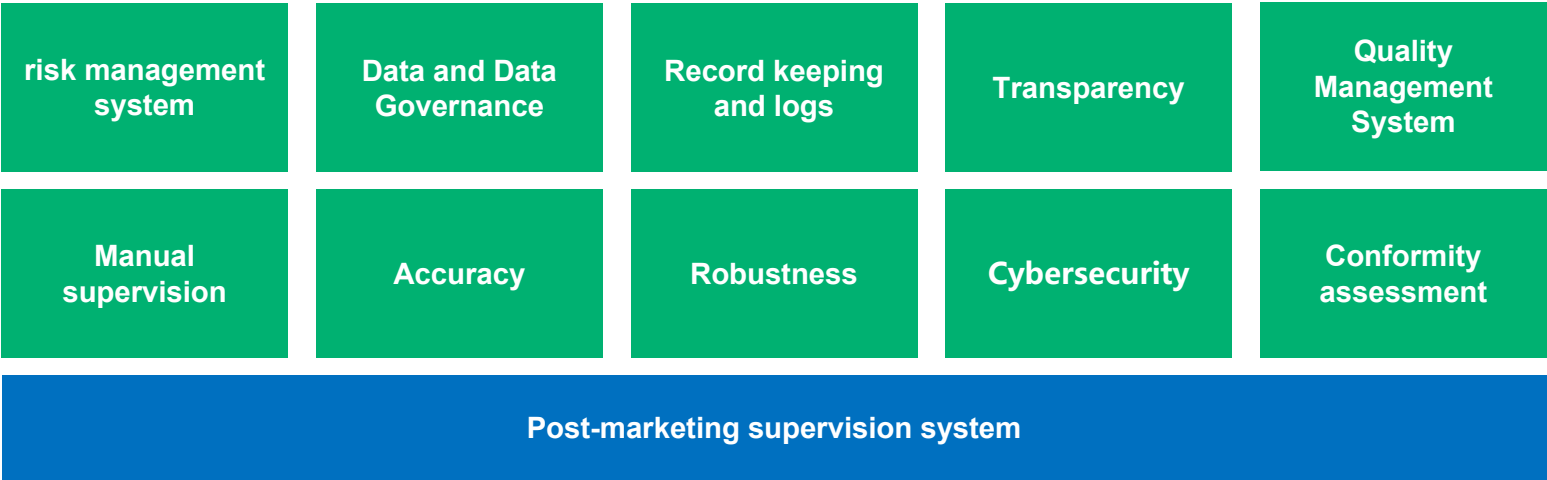
Every legal obligation produces testable evidence — the evaluation framework translates compliance requirements into measurable evidence criteria

Market Obligations and Application Timeline

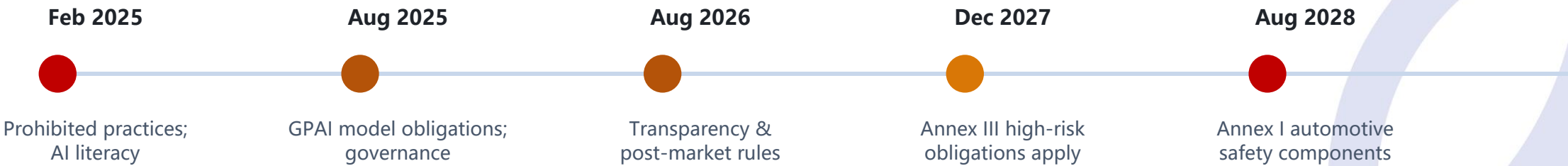
EU market access combines mandatory lifecycle controls with phased high-risk implementation

Mandatory requirements for the use of high-risk artificial intelligence systems in the EU market:

For cockpit AI, these duties span the lifecycle: risk/data governance, oversight and cybersecurity; post-market monitoring and conformity assessment.



PHASED APPLICATION TIMELINE



High-risk dates reflect the AI Omnibus, Regulation (EU) 2026/1744 (in force 27 Jul 2026). Automotive OEMs must complete high-risk compliance before Dec 2027.

Non-compliance carries both market-access risk and direct financial exposure for OEMs and AI providers

03

Trustworthy Evaluation System

A full-lifecycle evidence chain — building verifiable safety baselines at every integration point.

Evaluation Paradigm Shift: From Performance to Trustworthiness

Trustworthiness evaluation adds safety, resilience, oversight and lifecycle evidence to conventional performance metrics

Existing: Performance-Based Evaluation

What it measures:

- Accuracy & quality: WER, mAP, recognition rate
- Speed & stability: latency, throughput, response time
- Coverage & usability: scenarios and user experience
- Output: benchmark scores and feature comparisons

Key limitations:

- No safety-boundary or dangerous-command tests
- No adversarial or fault-injection evaluation
- No cross-domain interaction or liability trace
- No lifecycle monitoring or conformity evidence

PARADIGM
SHIFT

Integrated Trustworthiness Evaluation

What it adds (beyond performance):

- Safety: command boundaries, MRC, functional safety
- Robustness: adversarial and fault-injection tests
- Cybersecurity: prompt, tool and OTA attack tests
- Transparency: decision trace and AI disclosure
- Fairness: group parity and accessibility
- Controllability: interrupt, override, emergency stop
- Data Governance: provenance, logs and privacy

Key advantages:

- System-level, lifecycle evaluation
- Testable evidence for conformity assessment

Performance metrics show capability; trustworthiness evidence supports safe deployment

Each dimension pairs an EU AI Act duty with a concrete automotive failure mode, forming a traceable trust boundary.



**Art. 15
Accuracy**

Matches scenes and intent; correctly interprets commands and retrieval results.

**Art. 15
Robustness**

Stable under noise, weather, unseen inputs and faults.

**Art. 6/9
Safety**

Keeps actions within functional-safety bounds, degrading to minimal risk.

**Art. 15
Cybersecurity**

Resists adversarial inputs, prompt injection, tool hijacking and OTA attacks.

**Art. 11/13/50
Transparency**

Shows perceived decisions and rationale; decisions remain traceable.

**Art. 10
Fairness & Bias**

Delivers comparable service across accents, dialects, and user groups..

**Art. 14
Controllability**

Lets driver interrupt or stop, with visible authority limits.

**Art. 10/12
Data Governance**

Keeps training and log data lawful, traceable, and protected.

Mutually-Reinforcing Trust Boundary & Testability Principle

No dimension works alone. Accuracy must be constrained by safety; safety must be explainable with human oversight recorded. Every dimension is operationalised with evidence: dangerous-command and MRC tests for Safety; interruption latency and emergency-stop tests for Controllability; decision reconstruction and HMI for Transparency.

Trustworthiness Areas & SG17: From Evaluation to Security Standards

Q16/17 leads AI security; Q13/17 supplies the ITS/CAV context for automotive-specific evaluation

ITU-T SG17 STRUCTURE & AI SECURITY WORK

Q16/17 • AI Security

AI-agent requirements
Adversarial AI evaluation

Q13/17 • ITS/CAV Sec

Connected vehicle security
V2X and CAV scenarios

Q7/17 • Secure Apps

Application and service security
Secure invocation interfaces

Cross-cutting SG17

Architecture, IoT, identity
Privacy and assurance

Supporting Questions cover security architecture, IoT, identity, privacy and conformance methods.

EIGHT DIMENSIONS → SG17 QUESTION MAPPING

Accuracy	→ Q16/17 + Q13/17 · benchmarks
Robustness	→ Q16/17 + Q13/17 · adversarial/fault tests
Safety	→ Q13/17 + Q16/17 · boundaries and MRC
Cybersecurity	→ Q16/17 + Q13/17 · agents, tools and OTA
Transparency	→ Q16/17 · decision trace and disclosure
Fairness	→ Q16/17 · bias evaluation
Controllability	→ Q16/17 + Q13/17 · override and stop
Data Governance	→ Q16/17 + cross-cutting · lineage and logs

THE STANDARDIZATION GAP: AUTOMOTIVE AGENTIC AI

Existing SG17 AI standards

AI-agent, adversarial-AI and ITS/CAV security work already provide a strong foundation.



Automotive-specific gap

No unified automotive evaluation profile covers these trustworthiness areas across the full control loop.



Proposed contribution

Coordinated Recommendations linking methods, vehicle test profiles and conformity evidence.

SG17 has the right foundation; the missing layer is an automotive-specific security evaluation profile

The next two pages translate these areas into cockpit-focused standardization priorities

Cockpit Trustworthiness Areas → SG17 Security Evaluation (I)

**Benchmarks establish cabin baselines; adversarial and fault tests expose degradation;
Safety constrains vehicle functions; cybersecurity protects cockpit agents, APIs and updates**

Accuracy — EU AI Act Art. 15(1)(3)

SG17 RELEVANCE

Automotive AI needs shared multimodal datasets and thresholds for cockpit voice, driver intent and driving perception.

PROPOSED STANDARDIZATION FOCUS

Accuracy focus: Smart-cockpit multimodal benchmarks

Benchmark wake-word, speech/voiceprint, intent, gaze/gesture and screen-grounded responses across seats, noise and languages.

Robustness — EU AI Act Art. 15(4)

SG17 RELEVANCE

Robustness needs adversarial inputs, sensor-fusion faults and measurable degradation criteria for safety-critical use.

PROPOSED STANDARDIZATION FOCUS

Robustness focus: Cabin-condition resilience

Test cabin noise, occlusion, sensor loss and ASR/LLM perturbations, with fallback and recovery criteria.

Safety — EU AI Act Art. 6 / 9

SG17 RELEVANCE

Automotive AI must reject dangerous commands and reach a minimal-risk condition. Q13/17 supplies vehicle context; Q16/17 supplies AI-agent controls.

PROPOSED STANDARDIZATION FOCUS

Safety focus: Cockpit-agent action boundaries

Test unsafe voice commands, distracting actions, cross-domain control escalation and safe fallback to driver control.

Cybersecurity — EU AI Act Art. 15(5)

SG17 RELEVANCE

Security evaluation must cover prompt injection, tool/API misuse, model protection, data poisoning and OTA attacks.

PROPOSED STANDARDIZATION FOCUS

Cybersecurity focus: Cockpit agent & service security

Test voice/display prompt injection, unauthorized tool/API calls, app connectivity and OTA update attacks.

Cockpit Trustworthiness Areas → SG17 Security Evaluation (II)

Traceable cockpit decisions and equitable interaction support repeatable security evaluation
Driver override and trustworthy cabin data close the lifecycle evidence loop

Transparency — EU AI Act Art. 11 / 13 / 50

SG17 RELEVANCE

Evaluation should verify AI identity disclosure, decision reconstruction, HMI warnings and generated-content marking.

PROPOSED STANDARDIZATION FOCUS

Transparency focus: Explainable cockpit interaction

Standardize AI identity, action confirmation, source/rationale display, privacy indicators and generated-content marking.

Fairness — EU AI Act Art. 10(2)(f)

SG17 RELEVANCE

Security evaluation should detect unequal error rates or access across dialects, voiceprints, seats and user groups.

PROPOSED STANDARDIZATION FOCUS

Fairness focus: Inclusive cockpit interaction

Measure speech/voiceprint parity across accents, ages, genders and seats, plus accessible multimodal interaction.

Controllability — EU AI Act Art. 14

SG17 RELEVANCE

X.S-AIA provides agent oversight requirements; automotive profiles need measurable interruption, override and emergency-stop criteria.

PROPOSED STANDARDIZATION FOCUS

Controllability focus: Driver command & override

Measure cancel, interruption, confirmation, permission visibility, handover and emergency-stop response times.

Data Governance — EU AI Act Art. 10 / 12

SG17 RELEVANCE

Automotive AI needs verifiable lineage, poisoning defence, secure logs and personal-data protection.

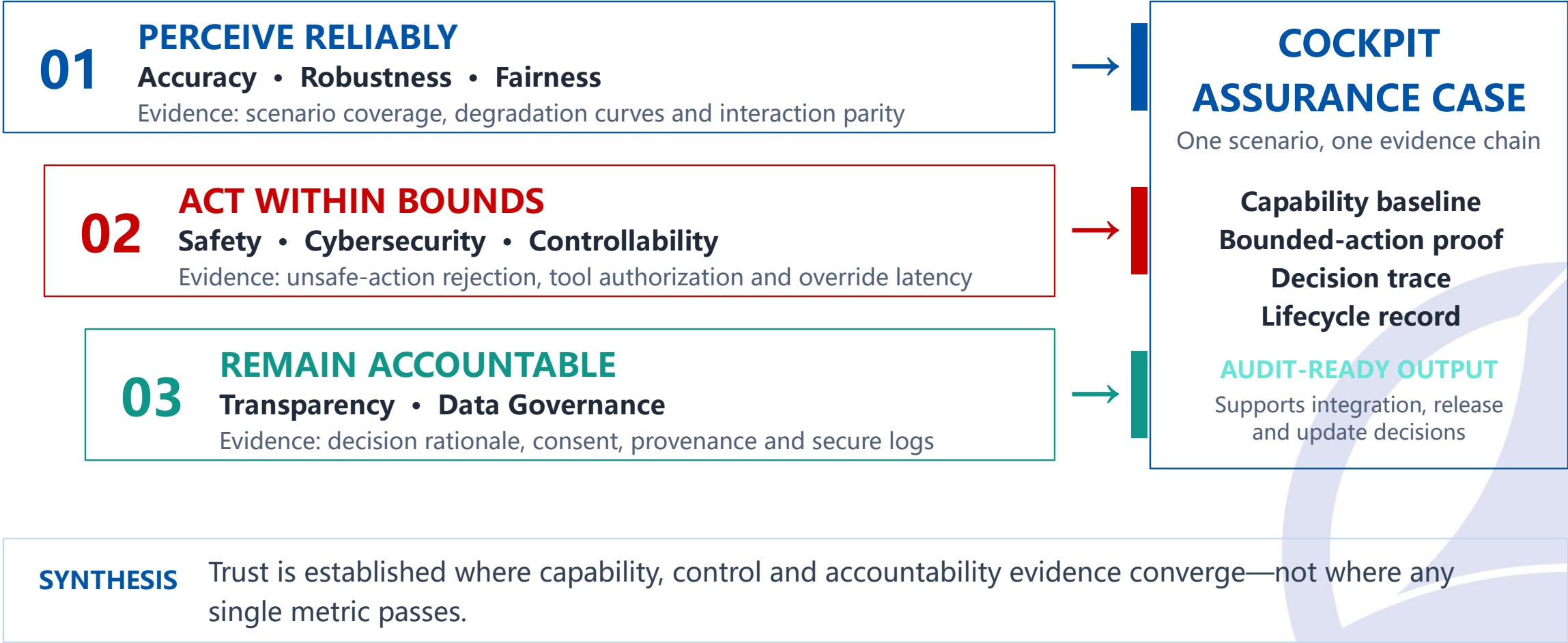
PROPOSED STANDARDIZATION FOCUS

Data focus: Cockpit-data lifecycle assurance

Define consent, minimization, retention, provenance and secure logs for cabin audio/video, biometrics, location and preferences.

Three Streams Build One Cockpit Assurance Case

Organize the eight trustworthiness areas by system role to create actionable evidence.





04

Summary & Outlook

From verifiable baselines to actionable deployment pathways

Key Takeaways & Future Directions

Four key takeaways and three future directions for establishing verifiable trust baselines in automotive agentic AI

1 Regulatory urgency is real and imminent

EU AI Act high-risk provisions apply from Dec 2027; automotive AI falls under Annex III and Annex I.

2 Performance-only evaluation is insufficient

Accuracy and speed metrics cannot capture safety boundaries, liability traceability, or adversarial robustness. Agentic AI requires multi-dimensional trustworthiness assessment.

3 Trustworthiness areas form an integrated framework

Accuracy, robustness, safety, cybersecurity, transparency, fairness, controllability and data governance map to EU AI Act Articles 9-15 and generate traceable conformity evidence.

4 A practical validation pathway is available

CAERI testing + TÜV SÜD certification provides internationally recognised conformity evidence and an actionable deployment pathway.

FUTURE DIRECTIONS

Standardisation

Contribute to ITU-T and ISO standards for agentic AI trustworthiness evaluation methodologies and metrics.

Cross-domain fusion

Develop evaluation methods for cross-domain agent interactions and end-to-end safety propagation.

Continuous assurance

Build post-market monitoring infrastructure with automated OTA re-certification triggers.

Trustworthiness is not a feature — it is a verifiable property that must be engineered, tested, and continuously assured



中国汽车工程研究院股份有限公司
China Automotive Engineering Research Institute Co., Ltd.

Thank You