Ethical and Operational Guidelines for Human-in-the-Loop ADAS and Driver Monitoring Systems

1. Purpose and Scope of the Guidelines

This document establishes a comprehensive set of **ethical and operational guidelines** for the responsible design, development, deployment, and governance of Artificial Intelligence (AI) in Human-in-the-Loop Advanced Driver Assistance Systems (ADAS) and driver monitoring applications.

The primary purpose of these guidelines is to translate Open and Responsible Research and Innovation (ORRI) principles into concrete, enforceable system requirements that can be operationalised within real-world engineering workflows. Rather than remaining at the level of abstract ethical commitments, the guidelines define actionable rules, constraints, and design obligations that directly inform system architecture, software development, human–machine interaction design, and organisational decision-making.

The document is intended to serve simultaneously as:

- an **internal reference framework** for AviSense engineering, product, and governance teams
- a **validated ORRI tool** to be disseminated through the REINFORCING platform
- and a **replicable model** for other SMEs and mobility actors developing AI-enabled, safety-relevant systems

1.1 Scope of Application

These guidelines apply to AI-based systems that influence, support, or condition driver perception, attention, decision-making, or behaviour, while explicitly preserving human authority and legal responsibility. The focus is on Human-in-the-Loop systems, where AI acts in an advisory or conditional support role rather than as an autonomous controller.

The scope explicitly includes, but is not limited to, the following domains:

Domain	Description
Semi-automated driving contexts	ADAS functions operating primarily at SAE Level >2, where the human driver remains continuously responsible
Driver condition monitoring	Detection and interpretation of fatigue, distraction, posture, gaze, or behavioural patterns
AI-assisted decision support	Warnings, alerts, recommendations, and conditional safeguards that influence driver behaviour

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the REA. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

Conditional interventions	Temporary, reversible system actions (e.g. feature limitation) requiring driver awareness and confirmation
Human–Machine Interfaces (HMI)	Visual, auditory, and AR-based interfaces that communicate AI outputs to the driver
Data governance	Processing of behavioural, contextual, vehicle, and (where applicable) biometric signals

The guidelines are technology-agnostic, meaning they apply regardless of specific sensor configurations, AI models, or hardware platforms, provided that the system falls within the above functional scope.

1.2 Explicit Exclusions

To avoid ambiguity, the following are out of scope for this document:

- Fully autonomous driving systems (SAE Level 4–5)
- Systems that remove or bypass human decision authority
- AI systems used exclusively for post-hoc analytics with no influence on live driving behaviour

These exclusions ensure that the guidelines remain focused, realistic, and enforceable within the context of TRUST-AI.

1.3 Regulatory and Normative Alignment

The guidelines are designed to ensure regulatory readiness and future-proof compliance, with explicit alignment to the following frameworks:

- **EU AI Act**. Classified as High-Risk AI systems under Annex III (road traffic safety)[1], with direct relevance to Articles 9 (risk management), 10 (data governance), 13 (transparency), 14 (human oversight), and 15 (accuracy and robustness).
- **General Data Protection Regulation (GDPR)**. In particular, principles of lawfulness, fairness, transparency, data minimisation, purpose limitation, and data subject rights (Articles 5–14)[2].
- **ORRI Framework**. Embedding anticipation, inclusion, reflexivity, and responsiveness throughout the system lifecycle, not as an external audit layer but as a design principle.

1.4 Validation and Evolution

These guidelines are not intended to be static. They constitute a living reference framework that will be:

[1] https://eur-lex.europa.eu/eli/reg/2024/1689/oj#d1e1403-1-1
[2] https://eur-lex.europa.eu/eli/reg/2016/679/oj#d1e1889-1-1

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the REA. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

- **Validated and stress-tested** through citizen and stakeholder engagement activities (Action 2),
- **Refined** based on societal feedback, ethical deliberation, and expert review,
- **Adopted** as a baseline for future system development at AviSense,
- **Reused and adapted** by external organisations via open dissemination.

Through this iterative process, the guidelines ensure that responsible AI in ADAS is not only compliant, but socially legitimate, technically grounded, and operationally sustainable.

1.5 How These Guidelines Are Used

These guidelines are intended to function as an operational reference throughout the full lifecycle of AI-enabled ADAS and driver monitoring systems, rather than as a static ethical statement. They are consulted at clearly defined stages of system design, development, deployment, and review, and are embedded into both technical and organizational decision-making processes. The guidelines are first applied during system architecture and feature design, where they inform decisions on task allocation between human and AI, selection of sensing modalities, data minimization strategies, and definition of human–machine interaction logic. At this stage, engineering teams are required to verify that any proposed ADAS or monitoring function complies with the foundational principles of human agency, proportionality, explainability, and privacy before implementation begins. They are subsequently consulted during feature approval and integration, particularly when introducing new AI-driven functions, modifying intervention thresholds, or deploying perception-based features such as AR visualizations or driver condition monitoring. Before approval, teams must demonstrate that override mechanisms, consent flows, and explanation strategies are explicitly designed and tested in line with these guidelines.

A dedicated application point is the HMI design and review process. The guidelines are used to assess whether system outputs, alerts, explanations, and visual cues are understandable, non-authoritative, and cognitively appropriate for drivers in real driving contexts. Any HMI element that influences driver behavior must be reviewed against the explainability and proportionality requirements defined in this document. Responsibility for applying and enforcing the guidelines is distributed across clearly defined roles. System and AI engineers are responsible for implementing the guidelines at the technical level, including data handling, decision logic, override functionality, and explainability mechanisms. An **Ethics and ORRI Focal Point** within AviSense oversees alignment between technical implementation and ethical requirements, supports teams in resolving ambiguities, and acts as the first escalation point for ethical concerns. Strategic oversight is provided by the **Internal ORRI Committee**, which reviews high-risk features, approves deviations where justified, and ensures alignment with regulatory and organizational commitments.

Compliance with the guidelines is verified through a combination of design-time and runtime checks. At design time, engineers use structured checklists derived from this document to confirm compliance before feature approval. During development and testing, key performance indicators (KPIs)—such as explainability coverage, override

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the REA. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

accessibility, response latency, and data retention limits—are monitored and documented. Periodic internal audits and ethical reviews, supported by evidence from system logs and validation exercises, assess whether implemented systems continue to operate in line with the guidelines over time.

2. Foundational Principles

All ethical and operational rules defined in this document are grounded in a coherent set of **five foundational principles**.

1. **Human agency and oversight** - The human driver remains the final decision-maker and legal authority at all times.
2. **Transparency and explainability** - AI system behavior must be understandable, inspectable, and explainable to non-expert users.
3. **Proportionality and safety** - System responses must be proportionate to assessed risk and avoid unnecessary restriction of user autonomy.
4. **Privacy and data minimization** - Data processing must be limited to what is strictly necessary for the defined function.
5. **Accountability and traceability** - Decisions, recommendations, and overrides must be traceable without enabling surveillance or punitive use.

These principles provide the normative backbone for the guidelines and serve as **design constraints** that shape technical decisions, system behavior, and organizational governance throughout the ADAS lifecycle. Rather than functioning as abstract values, each principle is interpreted in operational terms, ensuring that it can be embedded into engineering requirements, system architecture, and evaluation criteria.

2.1 Human Agency and Oversight

The human driver remains the **final decision-maker and legal authority** in all driving situations supported by ADAS. AI systems must never replace, obscure, or override human responsibility.

Operational interpretation

Human agency requires that the driver:

- retains continuous awareness of system activity
- has the ability to intervene, override, or disengage AI-supported functions
- is never forced into compliance with system recommendations

AI systems may influence perception and decision-making, but they must do so in a way that supports informed human judgment, not substitutes it.

Design implications

- All ADAS functions must be designed as **advisory or conditional**, not autonomous.
- Override mechanisms must be explicit, immediate, and intuitive.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the REA. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

- No irreversible action affecting vehicle behavior may occur without driver awareness.

This principle directly supports compliance with **EU AI Act Article 14 (Human Oversight)**.

2.2 Transparency and Explainability

AI system behavior must be **understandable, inspectable, and explainable** to non-expert users, including drivers without technical background.

Operational interpretation
Transparency is not limited to documentation; it must be experienced by the driver during system operation. Explainability refers to the system's ability to communicate:

- why a recommendation or alert was generated
- which observable factors contributed to it
- what level of confidence or uncertainty is associated with it

Design implications

- Every system output that may influence driver behavior must be accompanied by a human-readable explanation.
- Explanations must reference observable cues (e.g. driving behavior, environmental context), not internal model logic.
- Visual and AR-based interfaces must avoid creating an illusion of certainty when predictions are probabilistic.

This principle directly supports **EU AI Act Article 13 (Transparency and Provision of Information)**.

2.3 Proportionality and Safety

System responses must be **proportionate to the assessed level of risk** and must avoid unnecessary restriction of driver autonomy or excessive intervention.

Operational interpretation
Proportionality ensures that AI responses scale appropriately with risk severity and confidence. Minor or uncertain risk indicators should lead to gentle, informative alerts, while higher-confidence risks may justify stronger recommendations—but never disproportionate action.

Design implications

- Multi-level response strategies should be implemented (inform → warn → recommend).
- Conditional interventions must be reversible and time-limited.
- Safety mechanisms must be designed to support safe driving, not to enforce compliance.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the REA. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

This principle aligns safety engineering practices with ethical restraint, avoiding overreaction and automation bias.

2.4 Privacy and Data Minimization

Data processing must be strictly limited to what is **necessary, proportionate, and justified** for the specific ADAS function.

Operational interpretation
Privacy is embedded at the architectural level, not addressed retroactively. This requires careful consideration of:

- which data are collected
- how long they are retained
- where they are processed
- for what purpose

Design implications

- Behavioral and biometric data are processed only when essential.
- Raw biometric data are not stored by default.
- Processing should occur locally (edge) whenever possible.
- Cross-trip profiling and secondary data use are prohibited unless explicitly justified and consented.

This principle directly reflects **GDPR Articles 5 and 9** and supports trust in AI-enabled mobility systems.

2.5 Accountability and Traceability

System behavior must be **traceable and auditable**, while ensuring that traceability does not turn into surveillance or punitive monitoring of drivers.

Operational interpretation
Accountability requires that system decisions can be reviewed and evaluated after the fact, particularly in cases of disagreement, override, or incident. However, traceability must not undermine user trust or autonomy.
Design implications

- Key system events (alerts, recommendations, overrides) are logged in an abstracted, non-identifying form.
- Logs are used exclusively for:
    - ethical performance evaluation
    - system improvement
    - regulatory compliance
- Override actions are never treated as errors or violations.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the REA. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

This principle ensures alignment with both ethical governance and future conformity assessments under the EU AI Act.

2.6 Explicitly Prohibited Practices

To ensure that ethical principles are not only encouraged but **enforced through clear boundaries**, this document defines a set of explicitly prohibited practices. These prohibitions apply across all stages of design, development, deployment, and evaluation of Human-in-the-Loop ADAS and driver monitoring systems. The purpose of this section is to:

- Prevent ethically harmful system behaviors by design
- Reduce ambiguity in engineering and product decisions
- Support conformity with the EU AI Act, GDPR, and ORRI principles
- Demonstrate that TRUST-AI establishes enforceable limits, not only best-practice recommendations

Table 1: Prohibited Practices

Area	Prohibited Practice	Rationale
Explainability	Silent system interventions without user-facing explanation	Violates transparency and informed decision-making; increases automation bias
Explainability	Use of technical, model-centric explanations (e.g. scores, probabilities without context)	Not interpretable by non-expert users; undermines meaningful transparency
AR Interfaces	Photorealistic or highly realistic hazard overlays presented as factual certainty	Creates false certainty and over-trust; misrepresents system uncertainty
AR Interfaces	Visual dominance of AR cues over real-world perception	Risks cognitive capture and distraction; undermines situational awareness
Human Override	Delayed, hidden, or technically constrained disengagement mechanisms	Undermines human agency and legal responsibility
Human Override	Automatic re-engagement of ADAS after explicit driver override	Violates user intent and autonomy
Data Governance	Cross-trip behavioral or biometric profiling without explicit justification and consent	Violates GDPR principles and erodes trust
Data Governance	Long-term storage of raw biometric data by default	Disproportionate to purpose; increases privacy risk
Logging & Monitoring	Interpretation of override actions as driver errors or misuse	Encourages punitive surveillance and discourages responsible override
Logging & Monitoring	Use of ethical or safety logs for disciplinary, insurance, or enforcement purposes	Contradicts ORRI values and chills user trust

2.7 Mapping of TRUST-AI Guidelines to EU AI Act Requirements

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the REA. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

To ensure regulatory clarity, auditability, and alignment with forthcoming conformity assessments, the ethical and operational guidelines defined in this document are explicitly mapped to the relevant provisions of the **EU Artificial Intelligence Act**. This mapping demonstrates that the TRUST-AI guidelines are not abstract ethical commitments, but concrete operational rules that support compliance with **High-Risk AI system requirements**, as defined in **Annex III of the AI Act** for safety components in vehicles and driver monitoring systems.

The table below (Table 2) establishes a clear traceability link between guideline areas, applicable EU AI Act Articles, and the types of operational evidence generated within the TRUST-AI project. This structure enables future internal audits, external reviews, and regulatory assessments to verify compliance in a systematic and transparent manner.

Table 2: Guideline / EU AI Act Mapping

Guideline Area	Key Rules Defined in TRUST-AI	EU AI Act Article(s)	Operational Evidence Produced
Explainability & Transparency	Mandatory explanations for all alerts, warnings, and recommendations; explanations expressed in plain, non-technical language; uncertainty communicated where relevant	**Art. 13 – Transparency and Provision of Information**	HMI explanation screens; explanation logic specifications; system logs linking alerts to explanatory outputs
Human Oversight & Override	Driver retains final authority; override mechanisms always available, immediate, and intuitive; advisory vs intervention modes clearly communicated	**Art. 14 – Human Oversight**	Physical and digital override controls; override event logs; HMI mode indicators; internal validation reports
Data Governance & Privacy	Data minimization; no long-term biometric storage by default; prohibition of cross-trip profiling; purpose limitation	**Art. 10 – Data and Data Governance**	Data flow diagrams; data lifecycle documentation; consent flow records; anonymization and retention policies
Risk Management & Proportionality	Proportional system responses; no irreversible interventions without human confirmation; early identification of ethical risk hotspots	**Art. 9 – Risk Management System**	Ethical risk tables; mitigation plans; design-time risk assessments; stakeholder validation feedback
Robustness, Accuracy & Uncertainty Awareness	Explicit handling of uncertainty; confidence thresholds for perception outputs; avoidance of false certainty (especially in AR content)	**Art. 15 – Accuracy, Robustness and Cybersecurity**	Confidence indicators; uncertainty bounds in AI outputs; robustness testing results; performance monitoring logs

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the REA. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

2.8 Role of the Foundational Principles in TRUST-AI

This mapping is used as a **living compliance reference** throughout the TRUST-AI project. Each guideline defined in Sections 3–6 of this document can be traced to one or more AI Act obligations, ensuring that ethical design choices simultaneously support legal conformity. Table 2 also serves as a **bridge between technical documentation and regulatory language**, reducing ambiguity for engineers, ethics reviewers, and external stakeholders. It will be reused in:

- internal conformity self-assessments,
- stakeholder validation discussions (Action 2),
- the Internal ORRI Strategy and Ethics Charter (Action 4),
- and the evaluation framework and KPIs (Action 6).

3. Guidelines on Explainability

3.1 Explainability as a System Requirement

Within TRUST-AI, explainability is defined as a core system capability and a prerequisite for the ethical deployment of AI in Human-in-the-Loop ADAS and driver monitoring systems. It is not treated as an optional user-experience enhancement or a post-hoc documentation exercise, but as a functional requirement equivalent in importance to safety, robustness, and usability. Any ADAS function that influences driver's perception, attention, or behavior -whether through alerts, recommendations, visual cues, or conditional safeguards- must be capable of providing a clear, timely, and comprehensible rationale for its outputs. Explainability is essential to:

- Support informed human decision-making
- Prevent automation bias and over-reliance
- Maintain trust calibration between driver and system
- Enable accountability and auditability

From an ORRI perspective, explainability operationalizes the principles of transparency, reflexivity, and responsibility, ensuring that drivers are not subjected to opaque or inscrutable system behavior.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the REA. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

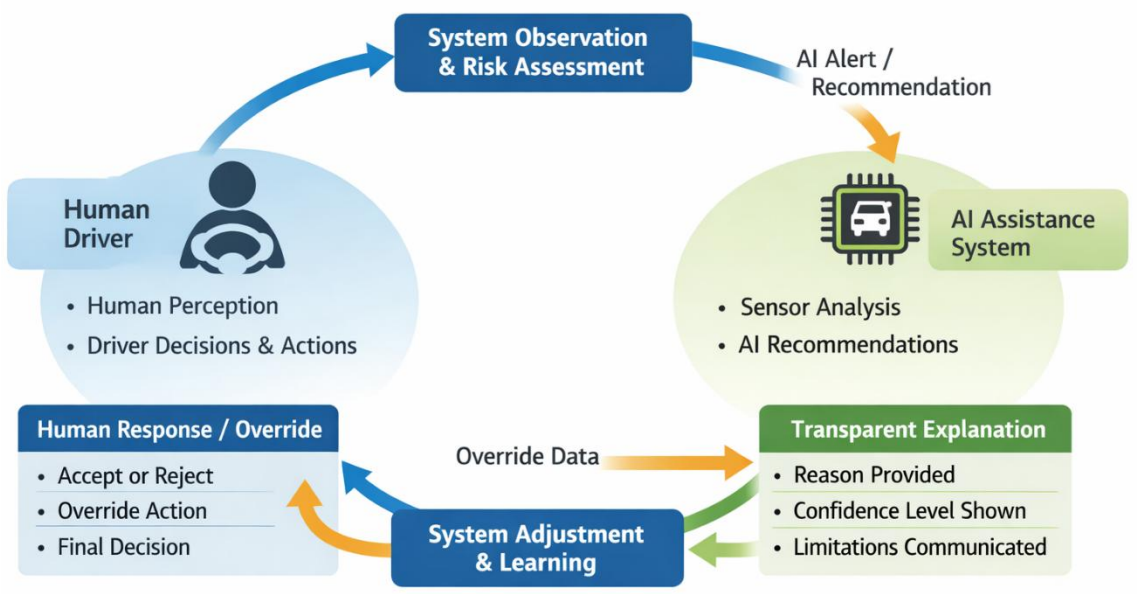

Figure 1: Explainable Human–AI Decision Loop

3.2 Operational Explainability Rules

To ensure consistency and enforceability, the following operational rules apply across all relevant ADAS functions. Every alert, warning, recommendation, or AR-based cue must be accompanied by a contextual explanation that is intelligible to a non-expert driver. Explanations must be expressed in plain, non-technical language and must be delivered through the same or a closely coupled HMI channel as the primary system output. Explanations must explicitly reference observable or experiential factors, such as: i) detected driving behavior (e.g. irregular steering, prolonged lane deviation), ii) environmental context (e.g. reduced visibility, dense traffic), iii) temporal conditions (e.g. extended driving duration without breaks).

Internal model metrics, abstract confidence scores, or references to algorithmic internals (e.g. “model probability = 0.82”) must not be exposed directly to the driver. Where technically feasible and relevant to risk perception, confidence levels or uncertainty indicators must be communicated. These indicators should be qualitative or visual (e.g. low / medium / high confidence, color-coded cues) rather than numeric, and must be designed to inform judgment without overwhelming attention. Explainability mechanisms must be integrated into the HMI in a way that minimizes cognitive load. Explanations should be concise, prioritised, and context-aware, avoiding excessive text or complex visual overlays that could distract the driver during critical driving moments.

3.3 Explainability in Advanced Interfaces (including AR)

For visual and AR-based interfaces, explainability carries additional ethical and safety implications. Visual cues must be designed to communicate uncertainty explicitly and avoid creating an illusion of objectivity or completeness. In particular:

- AR overlays must be abstract and symbolic rather than photorealistic.
- Visual elements should clearly indicate that content represents AI-inferred possibilities, not ground truth.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the REA. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

- Supplementary explanatory labels (e.g. “possible pedestrian detected behind vehicle”) must be displayed where appropriate.

This ensures that explainability extends beyond verbal explanation and is embedded directly into perceptual design choices.

3.4 Prohibited Explainability Practices

To prevent ethical and safety failures, the following practices are explicitly prohibited within TRUST-AI-aligned systems. Silent interventions or recommendations that influence driver behavior **without any accompanying explanation** are not permitted. Drivers must never be left to infer system intent implicitly. Explanations that rely on **overly technical, vague, or abstract terminology** that cannot reasonably be understood by a lay driver are prohibited. This includes references to AI models, scores, or computational processes that lack experiential meaning. Visual or AR representations that imply certainty, precision, or inevitability in situations where system confidence is low or moderate are also prohibited. Such representations risk inducing over-trust and automation bias, undermining human agency and safety.

3.5 Regulatory and ORRI Alignment

These explainability guidelines directly support:

- **EU AI Act Article 13**, by ensuring that users receive meaningful information about system operation,
- **EU AI Act Article 14**, by enabling effective human oversight,
- **ORRI principles**, by making AI behavior transparent, contestable, and socially accountable.

Explainability thus functions as a technical safeguard, an ethical obligation, and a regulatory enabler within TRUST-AI.

4. Guidelines on Human Override and Control Logic

4.1 Override as a Core Ethical and Safety Safeguard

Within TRUST-AI, human override is treated as a fundamental safeguard, not as an exceptional fallback mechanism. Override functionality operationalizes the principle of **human agency and oversight**, ensuring that AI systems supporting driving behavior remain subordinate to human judgment under all operational conditions. In Human-in-the-Loop ADAS, AI systems may influence perception, attention, or decision-making, but they must never remove or obscure the driver’s ability to intervene, disagree, or disengage. Override mechanisms therefore function as both a **safety control**, enabling immediate human intervention, and an **ethical control**, preventing automation bias, over-reliance, and loss of agency. Override design is particularly critical in semi-automated driving contexts, where ambiguity about responsibility or system authority can lead to unsafe behavior or misplaced trust.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the REA. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

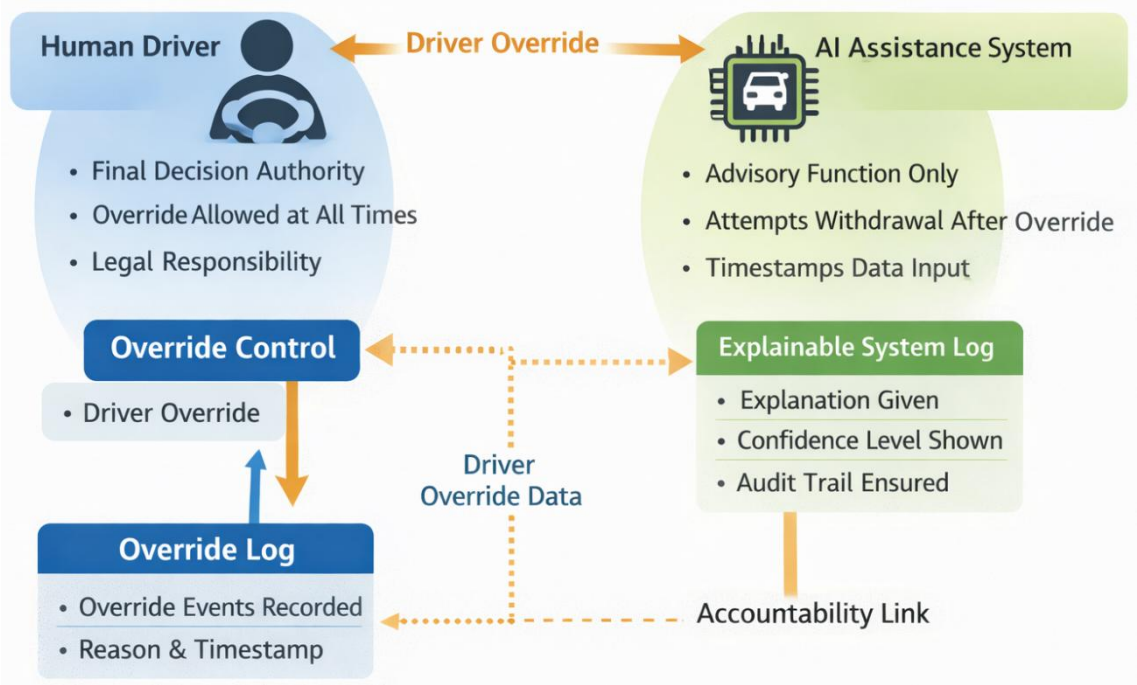

Figure 2: Override & Accountability Mechanism

4.2 Operational Rules for Override and Control

To ensure clarity, reliability, and usability, the following operational rules apply to all ADAS functions covered by these guidelines. Drivers must be able to override or disengage AI-supported functions through controls that are clear, immediate, and intuitive, without requiring complex sequences, hidden menus, or prolonged attention. Override controls must be accessible within the normal driving posture and must not require visual confirmation during critical moments. Override actions must never be:

- delayed by system logic,
- restricted based on contextual conditions,
- or disabled due to system confidence or internal thresholds.

The system must explicitly communicate its current operational mode at all times. In particular, drivers must be able to clearly distinguish between **advisory mode**, where the system provides information or recommendations only, and **conditional intervention mode**, where the system may temporarily modify or limit certain functions while still allowing full human control. Mode transitions must be communicated in real time through the HMI using unambiguous visual, auditory, or combined cues.

4.2 Override Logging, Accountability, and Ethical Use

Override actions may be logged, but only under strict ethical and governance constraints. Logging of override events is permitted solely for: i) safety analysis, ii) system improvement and debugging, iii) ethical performance evaluation within TRUST-AI, and iv) regulatory conformity assessment where required.
Override events must never be interpreted as system misuse, treated as user errors, or used for profiling, scoring, or penalizing drivers. To prevent surveillance or misuse,

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the REA. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

logged override data must be abstracted and anonymized wherever possible, excluding personally identifiable information and limiting context to what is strictly necessary for analysis. Logs must be time-bound and retained only for predefined, justified periods. From an ORRI perspective, override logging supports accountability without coercion, enabling learning and improvement while preserving trust.

4.4 Design Implications for ADAS and HMI

The requirement for meaningful override has direct implications for system and interface design:

- Override controls must be included in early system architecture decisions, not retrofitted.
- AR-based and visual interfaces must never obscure or dominate override options.
- System messaging must explicitly reinforce that AI recommendations are non-binding.

By embedding override as a first-class design element, systems align with **EU AI Act Article 14 (Human Oversight)** and strengthen user trust through visible respect for human authority.

4.5 Role of Override in TRUST-AI Governance

Within TRUST-AI, override behavior will be:

- discussed and validated with citizens and stakeholders during Action 2,
- reflected in internal ORRI Strategy and Ethics Charter (Action 4),
- evaluated through ethical KPIs such as override frequency, clarity, and user satisfaction (Action 6).

Override mechanisms thus function not only as technical controls, but as measurable indicators of responsible AI implementation.

5. Guidelines on Consent and User Awareness

5.1 Informed Consent by Design

Within TRUST-AI, consent is treated as a continuous, contextual process, not as a one-time administrative action performed during initial system setup. In Human-in-the-Loop ADAS and driver monitoring systems, consent must be understood as an ongoing relationship between the driver and the AI system, evolving with system state, functionality, and context of use. Drivers must be able to understand what data are being processed, for what purpose, and with what implications, at the moment that processing occurs—not only retrospectively through documentation. Consent therefore functions as a **design principle**, embedded directly into system architecture, HMI design, and operational workflows. This approach ensures that drivers remain informed, empowered, and capable of exercising meaningful control over AI-supported functions throughout system operation.

5.2 Operational Consent Rules

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the REA. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

All ADAS and driver monitoring functions covered by these guidelines must comply with the following operational consent requirements. Drivers must be clearly informed whenever monitoring, inference, or AI-assisted decision-support functions are active. This information must be provided through persistent, easily recognizable indicators integrated into the HMI and must not rely solely on user manuals or legal notices.
Consent mechanisms must be:

- Integrated into the initial system onboarding process.
- Accessible during normal system operation.
- Designed to allow drivers to review, modify, or withdraw consent without friction.

Where legally and technically permitted, drivers must be able to temporarily disable monitoring or inference functions—such as fatigue or distraction detection—without losing access to core vehicle functionality. Disabling such functions must not result in punitive system behavior, degraded vehicle control, or repeated coercive prompts. Any limitations imposed by legal or safety requirements must be communicated transparently and in plain language.

5.3 User Awareness and Transparency Requirements

User awareness extends beyond consent and includes real-time understanding of system behavior. To support this, systems must provide continuous transparency regarding their operational state. System status indicators must clearly and persistently communicate: a) which AI functions are currently active, b) whether the system is operating in advisory or conditional intervention mode, c) and whether data monitoring or inference is ongoing.
Any change in system behavior triggered by monitoring or AI inference—such as escalation of alerts, modification of recommendations, or temporary limitation of optional features—must be communicated to the driver in real time. Such communication must include a brief explanation of the reason for the change and the available driver options.
Transparency mechanisms must be designed to inform without distracting, using concise messaging, consistent visual language, and prioritization aligned with driving safety.

5.4 Data Protection and Regulatory Alignment

These consent and awareness guidelines directly support:

- **GDPR requirements** for lawful, fair, and transparent processing (Articles 5–7 and 12–14),
- **EU AI Act transparency obligations** (Article 13),
- and **EU AI Act human oversight provisions** (Article 14).

By embedding consent and awareness into system operation rather than treating them as legal formalities, TRUST-AI ensures that responsible AI deployment in mobility contexts is both legally compliant and socially legitimate.

5.5 Role of Consent and Awareness in TRUST-AI Validation

Consent and awareness mechanisms will be:

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the REA. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

- Tested and discussed with citizens and stakeholders during **Action 2**.
- Evaluated through user understanding and trust-related KPIs in **Action 6**.
- Incorporated into the Internal ORRI Strategy and Ethics Charter (**Action 4**).

This ensures that consent is not only technically implemented but also experienced as meaningful and respectful by real users.

6. Guidelines on Data Governance and Protection

6.1 Data Minimization and Purpose Limitation

All AI-enabled ADAS and driver monitoring functions must adhere to a strict **data minimization and purpose limitation regime**, ensuring that personal, behavioral, and biometric data are processed only, when necessary, only for a clearly defined function, and only for the duration required. For each ADAS function, the specific data inputs required must be explicitly documented at design time, including their purpose, sensitivity level, and retention constraints. No data may be collected "just in case" or for unspecified future use. Only data that are strictly necessary to deliver a defined safety or decision-support function may be processed. Where equivalent performance can be achieved using less sensitive data, the less intrusive option must be selected by default. Raw biometric data—such as facial imagery, gaze vectors, or posture signals—must not be stored by default. Where biometric signals are used, processing should focus on derived, task-specific features (e.g. fatigue indicators) rather than raw measurements. Cross-trip profiling, behavioral aggregation across journeys, or longitudinal driver characterization is explicitly prohibited, unless all of the following conditions are met:

i. A clear safety or regulatory justification exists
ii. Explicit and informed user consent is obtained
iii. The profiling scope is narrowly defined and documented
iv. A dedicated ethical risk assessment has been performed

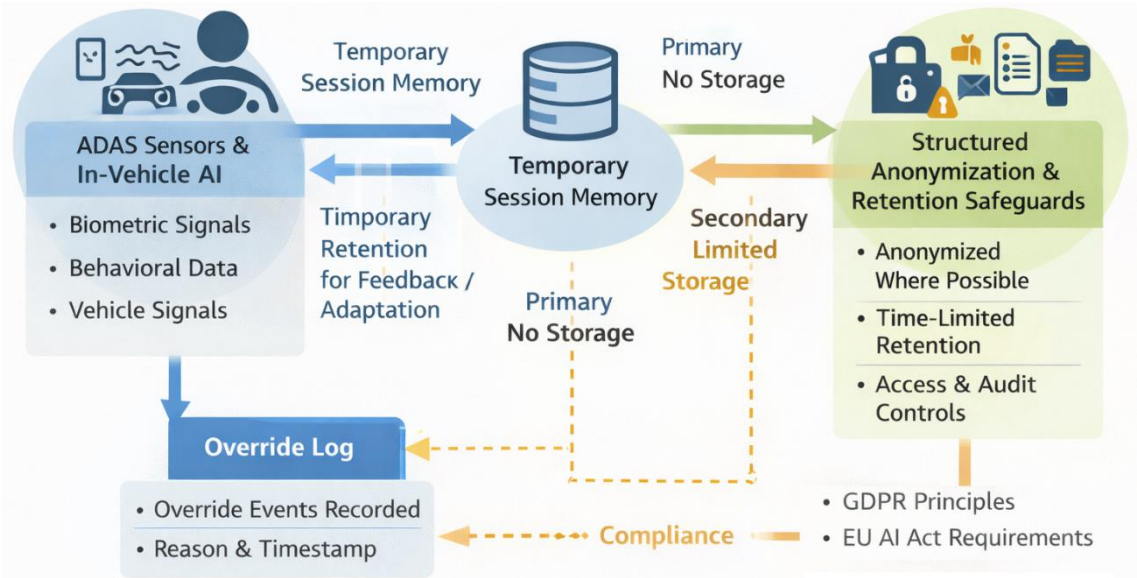

Figure 3: Data Lifecycle & Governance Architecture

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the REA. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

6.2 Storage, Retention, and Processing Architecture

Data processing should be designed to occur as close to the source as possible, favoring edge or in-vehicle processing over centralized storage. This reduces exposure risk, limits data propagation, and enhances driver trust. Temporary, session-based memory may be used to support: short-term system adaptation, real-time feedback loops, and post-event explainability during the same driving session.
Such session memory must be automatically cleared at the end of the trip unless retention is strictly required for safety verification or regulatory compliance. Long-term storage of raw signals—including video streams, biometric measurements, or detailed behavioral logs—is avoided by default. Where retention is mandated (e.g. for safety investigations or conformity assessment), stored data must be i) limited in scope, ii) time-bound with predefined deletion schedules, iii) secured using appropriate technical and organizational safeguards, iv) and subject to internal access control and audit mechanisms.

6.3 Data Governance Controls and Accountability

All data flows within ADAS and driver monitoring systems must be documented through data flow diagrams and governance registers, clearly indicating:

- data origin,
- processing purpose,
- retention duration,
- access rights,
- and deletion triggers.

Accountability for data governance must be clearly assigned within the organization, including defined roles for:

- system architects,
- data protection officers or ethics focal points,
- and responsible engineers.

Drivers must not be subjected to hidden secondary uses of their data, such as commercial profiling, insurance scoring, or behavioral monetization. Any deviation from the primary safety-related purpose constitutes a breach of these guidelines.

6.4 Compliance Alignment and Future-Proofing

These data governance guidelines are aligned with:

- **GDPR principles** of lawfulness, fairness, transparency, data minimization, purpose limitation, storage limitation, and integrity,
- **EU AI Act requirements** for training, validation, and operational data governance (Article 10),
- and ORRI commitments to responsibility, anticipation, and societal trust.

By embedding data governance at the architectural and procedural level, these guidelines support future **AI Act conformity assessments**, third-party audits, and institutional accountability.

6.5 Role of Data Governance in TRUST-AI Validation

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the REA. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

Data governance practices defined in this section will be:

- reviewed and stress-tested during citizen and stakeholder engagement activities (**Action 2**),
- evaluated through ethical performance KPIs (**Action 6**),
- and institutionalized through AviSense's Internal ORRI Strategy and Ethics Charter (**Action 4**).

This ensures that data protection is not only legally compliant, but also perceived as legitimate, proportionate, and trustworthy by users.

7. Validation Through Stakeholder Engagement

The ethical and operational guidelines defined in this document will not remain static or internally defined. They will be systematically validated, challenged, and refined through structured stakeholder engagement activities, ensuring that they reflect real societal expectations, user perceptions, and domain expertise rather than purely technical assumptions. Validation will take place through three complementary engagement mechanisms implemented under **Action 2** of the TRUST-AI project.

7.1 Citizen Foresight Workshops

A series of facilitated citizen foresight workshops will be conducted to explore how drivers and road users perceive AI-supported decision-making in ADAS contexts. These workshops will use scenario-based methods and "what-if" dilemmas derived directly from the guidelines (e.g. fatigue detection errors, AR-based hazard visualization uncertainty). Participants will be asked to: i) assess the acceptability of specific system behaviors, ii) identify thresholds where support becomes intrusive, iii) and articulate expectations regarding explanations, warnings, and overrides. nsights from these sessions will be used to refine intervention thresholds, escalation logic, and the tone, timing, and format of system explanations.

7.2 Ethical Risk → Safeguard Mapping

To ensure that ethical considerations are not treated abstractly, TRUST-AI explicitly links **identified ethical risks** to **concrete technical and organizational safeguards**. This mapping demonstrates how ORRI principles are translated into enforceable system design decisions and operational controls.

The Ethical Risk → Safeguard Mapping serves four purposes:

- It makes ethical risks **explicitly traceable** to mitigation measures
- It supports **EU AI Act risk management and conformity assessment**

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the REA. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

- It enables **systematic validation** during stakeholder engagement (Action 2)
- It provides a measurable basis for **ethical performance evaluation** (Action 6)

Table 3: Ethical Risk Mitigation Matrix

Ethical Risk	System Context	Safeguard Implemented	Relevant Guideline Section
Automation bias	Driver fatigue alerts and warnings	Advisory framing with explicit reinforcement of driver authority	§3 (Explainability), §4 (Override & Control Logic)
Loss of agency	AR-based situational awareness warnings	Always-available, immediate override and disable controls	§4 (Human Override and Control Logic)
Bias	Driver monitoring and behavioral analysis	Adaptive, context-aware thresholds and continuous calibration	§6 (Data Governance and Protection)
Opacity	AI perception and inference modules	Mandatory explanation layer using observable, non-technical cues	§3 (Guidelines on Explainability)

7.3 Deliberative Digital Panel (DDP)

In parallel, a Deliberative Digital Panel will engage a broader and more diverse group of citizens over an extended period. Through moderated online discussions, participants will review simplified representations of the guidelines and react to simulated ADAS scenarios. The DDP will allow for:

- iterative feedback across time,
- comparison of viewpoints between different user groups,
- and identification of recurring concerns or misunderstandings.

DDP outputs will directly inform adjustments to:

- explanation strategies,
- consent and transparency mechanisms,
- and the framing of human–AI responsibility boundaries.

7.4 Expert and Stakeholder Review Sessions

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the REA. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

Finally, targeted review sessions will be conducted with expert stakeholders, including legal experts, mobility professionals, ethics specialists, and system designers. These sessions will focus on feasibility, regulatory alignment, and unintended consequences. Experts will assess:

- consistency with EU AI Act obligations,
- coherence with safety and liability frameworks,
- and operational feasibility within real-world engineering constraints.

7.5 Feedback Integration and Traceability

All feedback collected through these mechanisms will be systematically documented and mapped to specific guideline sections. Changes will be logged with clear rationales, creating a transparent revision history. Validation outcomes will be explicitly referenced in:

- the final version of these guidelines (D1.2),
- the Internal ORRI Strategy (Action 4),
- and the ethical performance KPIs defined in Action 6.

Through this process, the guidelines evolve from a technical framework into a socially validated reference standard, ensuring legitimacy, accountability, and long-term trust.

8. Adoption and Use as Reference Framework

Following validation through citizen and stakeholder engagement, these ethical and operational guidelines will be formally adopted as a **living reference framework** within the TRUST-AI project and beyond. Their adoption is designed to ensure that ORRI principles are not only documented, but actively embedded into organizational practice, engineering workflows, and evaluation mechanisms.

8.1 Institutional Adoption at AviSense

AviSense will formally adopt these guidelines as a reference document governing the design and deployment of Human-in-the-Loop ADAS and driver monitoring systems. The guidelines will be integrated into internal development processes, serving as a mandatory reference during:

- system architecture definition,
- feature design reviews,

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the REA. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

- safety and ethics assessments,
- and internal validation milestones.

Compliance with the guidelines will be overseen by AviSense's internal ORRI governance structure, ensuring continuity beyond the project duration.

8.2 Integration with the Internal ORRI Strategy (Action 4)

The guidelines will directly inform the Internal ORRI Strategy and Ethics Charter developed under Action 4. They will provide the operational backbone for higher-level commitments, translating values such as transparency, accountability, and human oversight into concrete design rules and decision criteria.

This alignment ensures consistency between strategic intent and technical implementation, reducing the risk of ethical principles remaining abstract or symbolic.

8.3 Contribution to Evaluation and KPIs (Action 6)

The guidelines will act as a baseline against which ethical performance is evaluated. Specific rules and safeguards defined herein will be mapped to measurable indicators, including:

- explainability compliance rates,
- availability and use of override mechanisms,
- adherence to data minimization rules,
- and user understanding and trust indicators.

These KPIs will be assessed during and after system testing, providing evidence of responsible AI integration.

8.4 Publication and Reuse through REINFORCING

To support openness and replication, the validated guidelines will be published on the REINFORCING platform under an open license. They will be accompanied by explanatory notes, use-case examples, and references to EU regulatory frameworks, enabling other SMEs, NGOs, and public bodies to adapt them to their own contexts.

By functioning simultaneously as an internal governance tool and a publicly available ORRI resource, these guidelines contribute to broader ecosystem learning and responsible digital transformation across Europe.

9. Conclusion

This document establishes a **practical, enforceable, and context-aware framework** for the responsible design, development, and deployment of Artificial Intelligence in Human-in-the-Loop Advanced Driver Assistance Systems. Rather than treating ethics as an abstract or post-hoc concern, the guidelines translate ORRI principles into **concrete**

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the REA. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

operational rules that can be directly integrated into engineering workflows, system architectures, and organizational governance.

By addressing explainability, human override, consent, data governance, and accountability in a unified manner, the framework supports **safer, more transparent, and socially acceptable AI-enabled mobility systems**, while preserving human agency and legal responsibility. The emphasis on traceability, proportionality, and uncertainty-aware design responds directly to the ethical challenges raised by semi-automated driving and aligns with emerging regulatory expectations under the EU AI Act and GDPR.

Crucially, these guidelines are not intended as static compliance documentation. Through validation via citizen engagement and expert review, and through their integration into internal strategies and evaluation mechanisms, they function as a **living reference framework** that evolves alongside technological and societal expectations. Their publication through the REINFORCING platform further ensures transferability and reuse, contributing to broader ecosystem learning and capacity building.

In this way, the guidelines support the core ambition of TRUST-AI: to demonstrate how responsible AI in mobility can be operationalized in real-world settings—building trust not through promises, but through transparent design choices, accountable governance, and sustained institutional change.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the REA. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

10. Annex A – Stakeholder Validation Protocol

A.1 Purpose of the Validation Protocol

This annex defines the structured protocol through which the Ethical and Operational Guidelines for Human-in-the-Loop ADAS Systems are validated, refined, and formally adopted within the TRUST-AI project.

The protocol ensures that validation is:

- **Participatory**, involving citizens and affected stakeholders
- **Expert-informed**, integrating independent ethical and technical review
- **Traceable**, with documented feedback and change logs
- **Actionable**, resulting in concrete guideline revisions

Validation is treated as a **design control mechanism**, not a dissemination activity.

A.2 Who Participates in Validation

Validation is conducted through three complementary groups:

A.2.1 Citizens and End Users

Participants include drivers, mobility users, and members of the public recruited through:

- Citizen foresight workshops
- Deliberative Digital Panel (DDP)

Citizens focus on **perceived fairness, clarity, trust, and acceptability** of AI system behaviors.

A.2.2 Expert Stakeholders

Expert reviewers include:

- Mobility and transport safety experts
- AI ethics and governance specialists
- Legal and data protection professionals

Experts assess **regulatory alignment, ethical robustness, and technical feasibility**.

A.2.3 Moderation and Oversight

Validation activities are facilitated by InterMediaKT moderators and overseen by:

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the REA. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

- AviSense Ethics Focal Point
- TRUST-AI ORRI Committee

This ensures neutrality, methodological consistency, and accountability.

A.3 What Is Validated

Validation focuses on **specific, decision-relevant elements**, not abstract principles. These include:

1. **Risk Thresholds** - Acceptability of alert triggers, fatigue scores, and AR confidence levels.
2. **Explainability Mechanisms** - Clarity, usefulness, and non-misleading nature of system explanations.
3. **Override and Control Logic** - Visibility, accessibility, and perceived legitimacy of override mechanisms.
4. **Consent and Transparency Practices** - User awareness of monitoring functions and system status.
5. **Ethical Boundaries** - Identification of unacceptable system behaviors or "red lines".

A.4 How Feedback Is Collected and Recorded

Feedback is captured using structured, auditable instruments:

- Moderated discussion templates (workshops)
- Scenario-based questionnaires
- DDP written contributions and voting results
- Expert review forms with qualitative and quantitative inputs

All feedback is:

- Anonymized and aggregated
- Categorized by guideline section
- Logged in a centralized **Validation Register**

A.5 How Changes Are Tracked and Implemented

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the REA. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

Validation leads to documented changes through a **controlled revision process**:

1. Each feedback item is tagged as:
    - Accepted
    - Partially accepted
    - Rejected (with justification)
2. Accepted feedback results in:
    - Updated thresholds
    - Revised wording
    - Added safeguards or constraints
3. A **Before/After Change Log** is maintained, recording:
    - Original guideline version
    - Stakeholder input
    - Revised guideline text
    - Responsible role

This ensures full traceability from input to decision.

A.6 Validation Flow and Adoption Process

Figure A1 – Stakeholder Validation and Adoption Flow

Logical Flow

Draft Ethical & Operational Guidelines

↓

Citizen Foresight Workshops & DDP

↓

Structured Feedback Collection

↓

Expert Stakeholder Review

↓

Guideline Revision & Change Log

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the REA. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

↓

ORRI Committee Approval

↓

Formal Adoption & Publication

A.7 Output and Reuse

The validated guidelines are:

- Formally adopted within AviSense engineering workflows
- Integrated into the Internal ORRI Strategy (Action 4)
- Used as reference for evaluation KPIs (Action 6)
- Published on the REINFORCING platform as a reusable ORRI asset

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the REA. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union