1. Purpose and Scope of the Mapping Report

This Mapping Report constitutes the first foundational deliverable of Action 1 within the TRUST-AI project. Its primary purpose is to systematically document and analyse how Artificial Intelligence (AI) components interact with human drivers in the context of Advanced Driver Assistance Systems (ADAS), with particular emphasis on human-in-the-loop decision-making, data processing pathways, and ethically sensitive interaction points.

The report is designed to make the internal logic of ADAS systems transparent and technically interpretable, both for engineering teams and for non-technical stakeholders involved in later co-design and evaluation activities. By explicitly mapping how AI systems perceive the environment, generate recommendations or interventions, and interact with human judgement, the report addresses a critical gap often observed in AI-enabled mobility systems: the lack of traceability between automated system behaviour, human agency, and accountability.

The scope of the report is threefold. First, it aims to render AI-driven decision processes explicit, inspectable, and traceable by documenting perception pipelines, inference stages, confidence estimation, and decision outputs, as well as the points at which human drivers are expected to intervene, override, or validate system behaviour. This includes both routine operational scenarios and edge cases where system uncertainty or ambiguity is elevated.

Second, the report seeks to identify ethical, legal, and societal risk points that emerge during ADAS operation. These risks include, but are not limited to, automation bias, opacity of system reasoning, inappropriate reliance on AI outputs, biased sensing or inference, privacy risks related to behavioural or biometric data, and unclear allocation of responsibility between human and machine actors. Importantly, these risks are identified at an early design and development stage, before system behaviours become fixed or normalized.

Third, the Mapping Report establishes a shared technical and ethical baseline that will be used throughout the TRUST-AI project lifecycle. The findings documented here will directly inform the participatory foresight workshops and citizen panels conducted in Action 2, where mapped decision points and risk areas will be exposed to external scrutiny and societal reflection. In addition, the mapped elements will serve as reference inputs for the development of ORRI-aligned evaluation indicators and key performance metrics in Action 6.

It is important to note that this report does not aim to prescribe final ethical solutions or policy decisions. Instead, it deliberately focuses on systematically identifying where ethical choices, trade-offs, and responsibilities arise within ADAS systems. By doing so, it creates the necessary conditions for informed co-design, responsible governance, and evidence-based ethical integration in subsequent project actions.

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. System Overview: Human-in-the-Loop ADAS Context

ADASs operate within a socio-technical driving environment in which responsibility is shared, but not equally distributed, between human drivers and automated systems. While ADAS technologies increasingly perform perception, interpretation, and decision-support functions, the human driver remains the legally accountable actor under current European traffic and liability frameworks. This asymmetry between growing technical autonomy and persistent human responsibility makes the Human-in-the-Loop (HITL) paradigm central to ethical and regulatory discussions around AI-enabled mobility. In this context, ADAS systems do not replace the driver but continuously interact with human perception, judgement, and action. AI modules observe the environment and the driver's state, generate recommendations or warnings, and in some cases initiate temporary interventions. The driver, in turn, is expected to interpret these outputs, decide whether to trust them, and retain ultimate control over the vehicle. This dynamic interaction creates a continuous feedback loop between human and machine, where timing, clarity, confidence, and transparency are critical for safety and trust.

The purpose of this section is to describe the operational context in which Human-in-the-Loop ADAS functions are deployed and to clarify the roles played by both AI systems and human drivers. By doing so, it establishes the background against which data flows, decision chains, and ethical risk points are later mapped in detail.

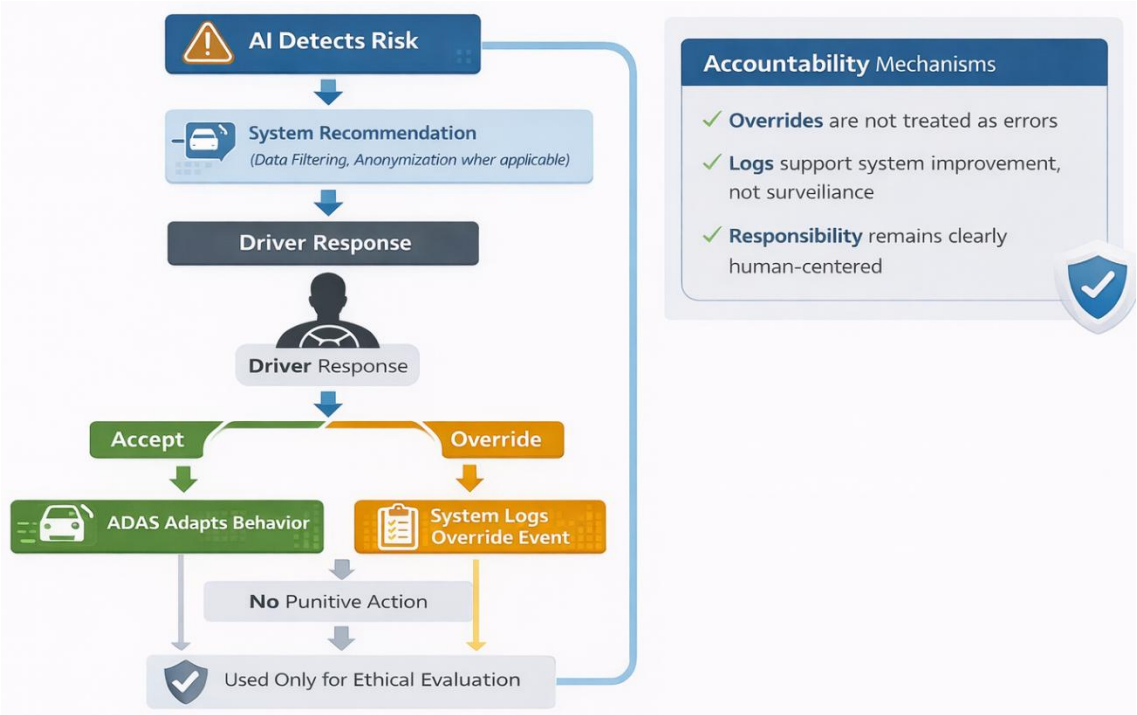

Figure 1: Override, escalation, and accountability mechanisms.

2.1 Objectives of the Human-in-the-Loop ADAS Mapping

The primary objective of mapping the Human-in-the-Loop ADAS context is to clearly delineate how decision authority, control, and responsibility are distributed between the driver and the AI system across different driving situations. This involves identifying

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

when AI systems merely provide information, when they influence driver behaviour through warnings or nudges, and when they actively intervene in vehicle operation.
Within the scope of this mapping, particular attention is given to ADAS functions that directly affect driver awareness, autonomy, and safety. These include driver condition monitoring functions that assess indicators such as fatigue, distraction, or posture; hazard detection and alerting mechanisms that identify risks in the vehicle's surroundings; AI-assisted interventions that may restrict or influence driver actions under specific conditions; and explicit driver override or system disengagement mechanisms that allow the human to reclaim control.

The mapping aims to clarify how these functions interact in real time, how system confidence and uncertainty are communicated to the driver, and how transitions between automated support and human control are handled. By explicitly documenting these interactions, the project seeks to expose potential points of friction, misunderstanding, or over-reliance that could compromise safety or undermine trust.
Ultimately, this section supports the broader objectives of TRUST-AI by framing ADAS not as isolated technical components, but as embedded socio-technical systems where ethical considerations emerge precisely at the interface between human agency and algorithmic decision-making.

3. Human–AI Interaction Mapping

This section describes how human drivers and AI-enabled ADAS components interact throughout the driving task, with particular emphasis on roles, responsibilities, and transitions of control. The objective is to make explicit how agency is distributed across the system and where ethical, legal, and safety-relevant decisions are effectively shaped by AI behaviour.

3.1 Roles and Responsibilities within the Human–AI System

Within a Human-in-the-Loop ADAS context, responsibility is distributed across multiple actors, but not equally. The human driver remains the primary decision-maker and the legally accountable controller of the vehicle. This responsibility persists even when AI systems actively support perception, assessment, or intervention. As such, the driver is expected to remain attentive, interpret system outputs, and intervene when necessary.
The ADAS AI operates as a supportive decision agent rather than an autonomous controller. Its role is to continuously process sensory inputs, assess risk conditions, and provide alerts, recommendations, or limited, predefined interventions. While the AI does not possess legal responsibility, it exerts significant influence over driver behaviour by shaping situational awareness, prioritising risks, and framing available choices. This influence introduces ethical considerations related to trust, automation bias, and transparency.
The vehicle control system acts as the execution layer of the overall architecture. It translates driver inputs and, where permitted, ADAS-generated commands into physical vehicle actions such as braking, steering resistance, or system disengagement. Although

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

technically deterministic, this layer becomes ethically relevant insofar as it enacts decisions originating from AI assessments that the driver may not fully understand.

Actor	Role
Driver	Primary decision-maker and legal controller
ADAS AI	Supportive decision agent providing alerts, recommendations, or limited interventions
Vehicle Control System	Executes physical actions based on driver input and/or ADAS recommendations

Crucially, the AI system does not replace human agency. Instead, it mediates decision-making by interpreting complex data streams and presenting simplified judgments to the driver. This mediation is where many ethical risks emerge, particularly if system logic, confidence levels, or limitations are insufficiently communicated.

3.2 Phases of Human–AI Interaction

The interaction between the driver and the ADAS AI can be understood as a sequence of four interconnected phases that repeat continuously throughout the driving task. Mapping these phases helps identify where ethical risks and responsibility shifts occur.

Phase 1 - Sensing and Perception. During this phase, the system collects data from multiple sources, including environmental sensors (such as cameras or radar), vehicle telemetry, and, where applicable, behavioural or biometric indicators related to driver condition. AI models process this data to construct an internal representation of the driving context and the driver's state. At this stage, the human driver plays a passive role, while the AI is actively interpreting inputs. Ethical concerns at this phase relate primarily to data minimisation, consent, and potential bias in sensing or inference.

Phase 2 - Assessment and Decision Support. Here, the AI evaluates perceived conditions against predefined thresholds, rules, or learned models to determine whether a situation requires attention. The system may choose to simply inform the driver, issue a warning, or recommend a specific action. The driver becomes an informed decision-maker, expected to understand and evaluate the system's output. The AI's role is advisory, but its framing of risk can strongly influence human judgement. Transparency, explainability, and clarity are critical at this stage to avoid over-reliance or confusion.

Phase 3 - Conditional Intervention. In certain predefined scenarios—typically related to safety—the system may temporarily intervene by limiting vehicle functionality, requesting explicit driver confirmation, or triggering safety mechanisms. Although such interventions are designed to reduce risk, they raise important ethical questions about proportionality, loss of autonomy, and appropriate thresholds for action. The driver is required to confirm or override the system, while the AI shifts from advisory to conditionally active.

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

Phase 4 - Override and Recovery. At any point, the driver must retain the ability to override or disengage the ADAS. When this occurs, full control is returned to the human, and the AI transitions to a passive or fallback role. This phase is essential for maintaining human authority and legal accountability. Poorly designed override mechanisms, delayed responses, or ambiguous system states can significantly undermine trust and safety.

By structuring Human–AI interaction into these phases, the mapping clarifies not only technical workflows but also where ethical responsibility, trust, and risk are negotiated in practice. This phased model provides the foundation for identifying ethical risk points and for aligning ADAS design with ORRI principles and EU AI Act requirements in subsequent sections of the report.

4. Data Flow Mapping

This section describes how data is collected, processed, stored, and governed within Human-in-the-Loop ADAS functions. The objective is to make data-related decisions explicit and traceable, given that data practices are a primary source of ethical, legal, and societal risk in AI-enabled mobility systems. Particular attention is paid to proportionality, consent, transparency, and accountability, in line with ORRI principles, GDPR obligations, and the EU AI Act.

4.1 Categories of Data Processed

ADAS functions rely on the combination of multiple data categories to support perception, risk assessment, and decision support.

Data Category	Examples
Behavioral data	Steering patterns, reaction time
Contextual data	Road conditions, traffic
Vehicle data	Speed, braking
Biometric data (if applicable)	Fatigue indicators, gaze direction

Behavioural data captures how the driver interacts with the vehicle, such as steering patterns, braking behaviour, reaction times, or lane-keeping performance. This data is used to infer driver attention, responsiveness, or potential impairment, but may also introduce bias if behavioural baselines are not representative.

Contextual data describes the external driving environment, including road geometry, traffic density, weather conditions, and the presence of vulnerable road users. This data is typically derived from onboard perception systems and, where available, cooperative sources such as infrastructure or nearby vehicles.

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

Vehicle data includes operational parameters such as speed, acceleration, braking force, and system state indicators. These signals are essential for both real-time safety functions and post-event analysis.

In certain configurations, biometric data may also be processed. This can include indicators related to fatigue, gaze direction, head pose, or posture. Given the sensitive nature of such data, its use is strictly limited to clearly defined safety purposes and subject to enhanced safeguards.

4.2 Data Processing and Lifecycle Flow

Data processing within the ADAS follows a layered and time-bound lifecycle. Data is first acquired through onboard sensors and, where applicable, external sources such as connected infrastructure or cooperative vehicle systems. This acquisition occurs continuously but selectively, based on system activation states.

Processing is primarily performed in real time at the edge, within the vehicle, to support immediate decision-making and to minimise latency and data exposure. AI models analyse incoming signals to assess driving context and potential risk conditions.

Temporary storage is used only for short-term system feedback, diagnostics, and safety logging. These logs enable traceability of system actions, support post-incident review, and contribute to quality assurance and ethical auditing.

As a core design principle, raw biometric data is not retained long-term. Persistent storage of such data is avoided unless explicitly required for safety validation or regulatory compliance, and only when clear user consent has been obtained. Where storage is necessary, data is minimised, pseudonymised, and protected by strict access controls.

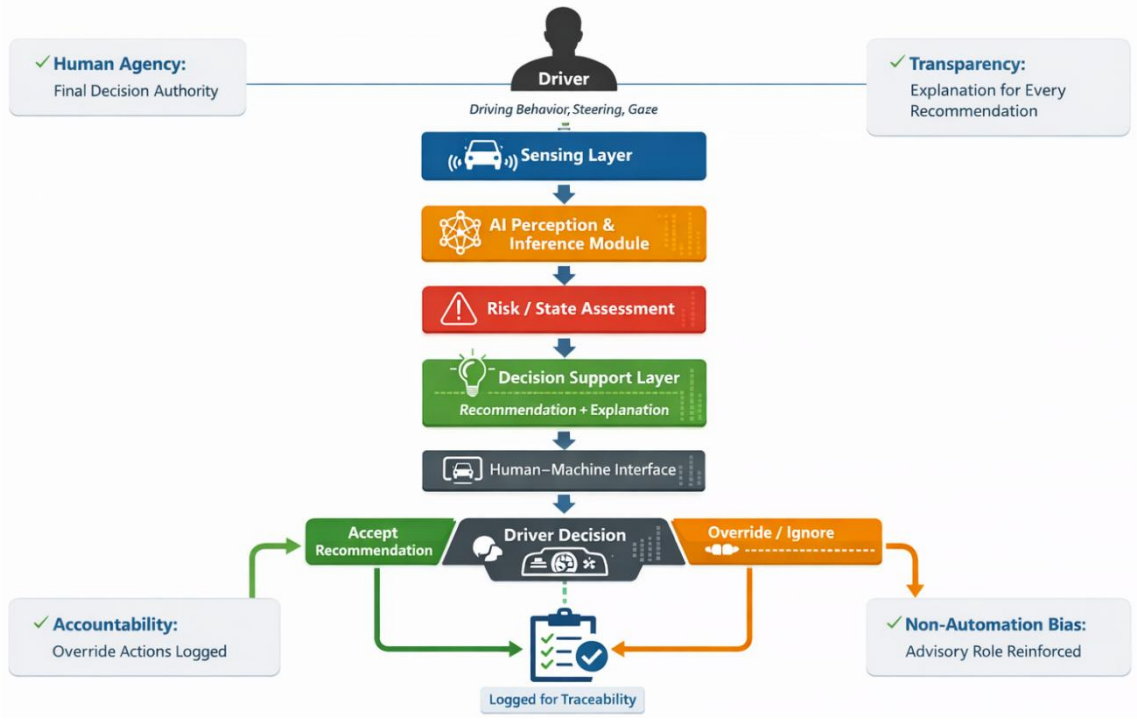

Figure 2: Logical Human-AI-in-the-Loop flow for ADAS decision support.

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

4.3 Consent Transparency, and Traceability Mechanisms

Transparency and user agency are integral to the data governance model. Drivers are clearly informed when monitoring or data-driven assistance functions are active, both during system onboarding and at runtime through interface cues. Consent mechanisms are embedded into initial system configuration and can be revisited or revoked by the user. All significant system actions related to data-driven decisions are logged in a structured and auditable manner. These logs do not serve surveillance purposes but enable traceability, accountability, and later ethical or technical review. They support internal evaluation processes and, where appropriate, third-party assessment of compliance with ORRI principles and regulatory requirements.

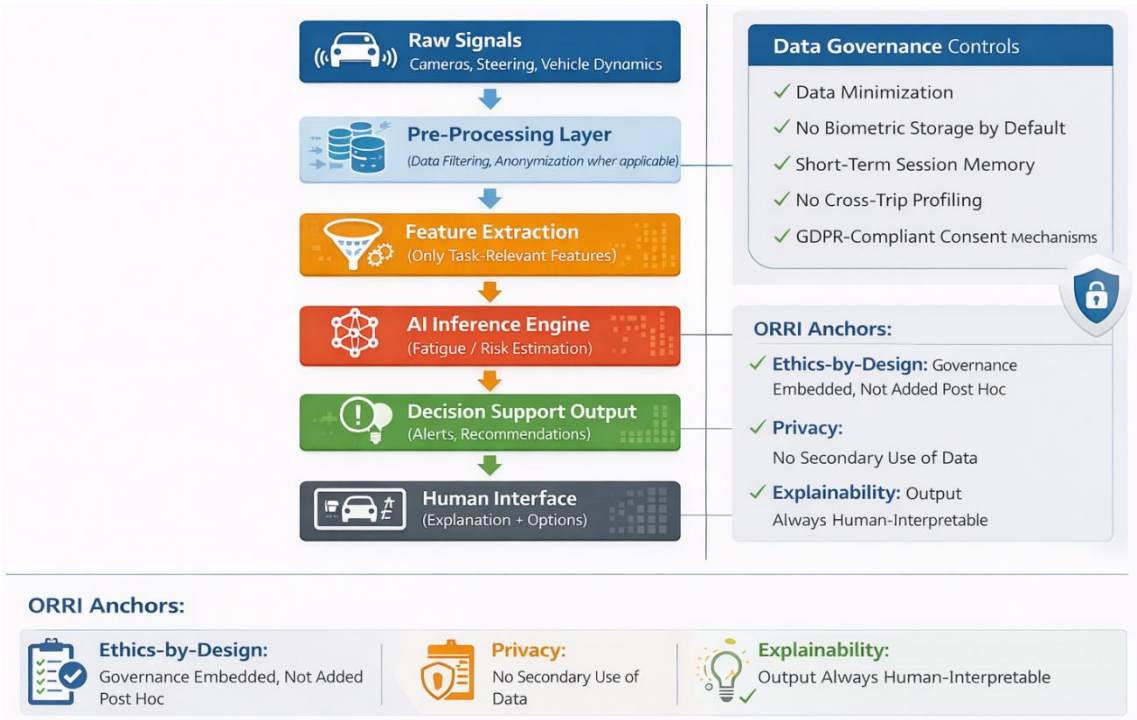

Figure 3: Data lifecycle and governance for Human-in-the-Loop ADAS.

5. Ethical Risk Identification

This section identifies and contextualizes the main ethical risks emerging from Human–AI interaction and data processing within ADAS. Rather than treating ethics as an abstract overlay, the mapping explicitly connects ethical risks to concrete interaction points, system behaviours, and design choices identified in earlier sections. The purpose is not to pre-emptively resolve these risks, but to make them visible, discussable, and actionable within subsequent co-design and evaluation phases.

5.1 Core Ethical Risk Categories

Risk Category	Description

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

Bias	Incorrect interpretation for certain user groups
Opacity	Lack of explainability of AI decisions
Automation bias	Over-trust in system recommendations
Inequality	Disproportionate false positives
Loss of agency	Excessive restriction of driver control

One of the primary risks concerns **bias**, particularly the risk that AI models may misinterpret behaviour or physiological signals for certain user groups. Variations related to age, physical ability, cultural driving styles, or atypical behaviour patterns may lead to systematic misclassification, especially in driver monitoring functions. If unaddressed, such bias can result in uneven system performance and unfair treatment of specific populations.

A closely related concern is **opacity**, where system decisions or interventions are not sufficiently explainable to the driver. When an ADAS system issues a warning, restricts functionality, or requests driver confirmation without a clear and understandable rationale, trust may be undermined and situational awareness reduced. Opacity also limits the ability of users and developers to challenge or improve system behaviour.

Automation bias represents another critical risk. Drivers may develop an over-reliance on system recommendations, gradually deferring judgement even in situations where human intuition or contextual understanding would be more appropriate. This risk is particularly pronounced in systems that operate reliably most of the time, creating a false sense of infallibility.

The mapping also highlights risks related to **inequality**, especially when false positives or conservative safety thresholds disproportionately affect certain users. For example, repeated unjustified interventions may discourage use of the system altogether or create cumulative disadvantages for specific groups.

Finally, there is the risk of **loss of human agency**, where excessive or poorly designed interventions constrain driver autonomy beyond what is necessary for safety. Even well-intentioned restrictions can become ethically problematic if they are perceived as coercive, non-negotiable, or insufficiently transparent.

5.2 Indicative Ethical Risk Hotspots

Several concrete risk hotspots emerge from the current ADAS configurations. One example is false detection of driver fatigue or distraction, which may trigger warnings or functional limitations without sufficient contextual justification. Another hotspot involves system-initiated restrictions that lack clear, real-time explanation, leaving drivers uncertain about why control has been limited.

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

Delayed, ambiguous, or poorly signposted override mechanisms also pose ethical concerns, particularly in time-critical situations where drivers must quickly regain full control. Additional risk points may arise as system complexity increases, especially when multiple AI modules interact or when cooperative data sources are introduced.

These identified risks are intentionally preliminary. They will be systematically validated, prioritized, and refined through citizen panels, foresight workshops, and stakeholder engagement activities under Action 2. This participatory validation will ensure that ethical risk prioritization reflects not only technical assessments, but also societal expectations, lived experience, and public values.

6. Traceability to Subsequent Actions

This Mapping Report constitutes a foundational artefact for the overall implementation of TRUST-AI and is explicitly designed to feed into subsequent actions of the project in a traceable and structured manner. Rather than remaining a static technical document, the mapping establishes a shared reference point that ensures continuity between analysis, participation, governance design, and evaluation.

First, the findings of this report directly inform **Action 2**, where citizen foresight workshops and the Deliberative Digital Panel will be conducted. The identified Human–AI interaction points, data flows, and ethical risk hotspots will be translated into concrete discussion prompts, scenarios, and dilemmas presented to citizens and stakeholders. This ensures that participatory engagement is grounded in real system behaviours rather than abstract ethical debate, allowing participants to meaningfully reflect on trade-offs, acceptable boundaries, and expectations for AI-supported driving.

Second, the mapping provides the technical and ethical baseline for **Action 4**, which focuses on the development of AviSense's internal ORRI Strategy and Ethics & Safety Charter. The documented interaction phases, risk categories, and governance touchpoints will be used to define internal policies, accountability roles, escalation procedures, and "red-line" thresholds for high-risk ADAS functionalities. In this way, insights from Action 1 are institutionalised into formal governance structures rather than remaining at the level of analysis.

Finally, this report establishes the reference framework for **Action 6**, where ORRI-aligned evaluation tools and key performance indicators will be defined. The mapped risks, decision chains, and transparency mechanisms will be translated into measurable indicators related to explainability, human agency, fairness, and trust. This ensures that ethical performance is assessed longitudinally and consistently, based on clearly documented system characteristics rather than post hoc perceptions alone.

Through this structured traceability, the Mapping Report functions as a living backbone of the TRUST-AI project, ensuring coherence between technical analysis, societal engagement, institutional change, and impact evaluation.

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

7. Concrete ADAS Use Case Walkthrough (Driver fatigue detection with Human-in-the-Loop override)

This section presents a concrete ADAS use case to demonstrate how human–AI interaction, data flows, ethical risks, and override mechanisms materialize in practice. The purpose of this walkthrough is not to evaluate system performance, but to illustrate how ORRI principles are translated into operational design decisions within a real ADAS function.

7.1 Use Case Description

The selected use case concerns an AI-based driver fatigue detection function operating in a semi-automated driving context. The system combines behavioral analysis, temporal driving patterns, and—where explicitly enabled—optional biometric indicators to assess early signs of driver fatigue. Its primary objective is to reduce accident risk by providing timely, explainable warnings while preserving driver autonomy, responsibility, and legal control.

The operational context is a privately owned vehicle equipped with Level >=3 ADAS capabilities. The human driver remains fully responsible for vehicle operation at all times, while the system functions strictly as a decision-support mechanism rather than an autonomous controller.

7.2 Step-by-Step System Operation

During the sensing phase (data acquisition), the system continuously collects a limited set of signals relevant to fatigue inference. These include behavioral indicators such as steering micro-corrections and lane-keeping variability, temporal patterns such as continuous driving duration, and—only when explicitly activated—optional biometric cues such as eye closure rate or postural changes. In line with ORRI and GDPR principles, data minimization is applied so that only information strictly necessary for fatigue assessment is processed.

In the interpretation phase (AI assessment), an AI model evaluates incoming data against adaptive thresholds. These thresholds are derived from contextual parameters, such as road type and speed, as well as short-term individual driving baselines established during the current trip. The system generates a fatigue likelihood score accompanied by confidence bounds. At this stage, the AI operates in a purely advisory capacity, without triggering any action that affects vehicle behavior or driver control.

When the fatigue likelihood exceeds a predefined threshold, the system enters the alert phase (decision support). The driver receives a visual and auditory warning, accompanied by a brief explanation describing why the alert was triggered—for example, irregular steering behavior observed over a defined time window. This explanation layer is mandatory and serves to mitigate automation bias by ensuring that the driver understands the system's reasoning rather than perceiving it as an opaque judgment.

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

If fatigue indicators persist, the system may transition into a safeguard phase (conditional intervention). In this phase, the system can recommend a rest break and, where legally permitted and technically appropriate, temporarily limit optional assistance features such as cruise assist. Importantly, no irreversible or safety-critical action is taken without explicit driver awareness, and the system remains subordinate to human authority.

At all times, the driver retains the ability to override system recommendations. Through clear and accessible human–machine interface controls, the driver can acknowledge the alert, explicitly override it, or disable the fatigue detection function for the current trip where regulations allow. All override actions are logged for accountability and system improvement purposes. This design ensures that human agency and final authority are preserved throughout the interaction.

7.3 Ethical Risk Analysis for the Use Case

This use case highlights several ethical risks inherent to fatigue detection systems. These include the potential for bias, such as false positives affecting certain driving styles or physical characteristics; opacity, where drivers may not understand why alerts are issued; over-reliance, where drivers defer judgment excessively to the system; and loss of agency if restrictions are perceived as coercive. Each of these risks is addressed through concrete mitigation measures, including adaptive thresholds, mandatory explanations, reinforcement of the advisory nature of alerts, and explicit override mechanisms.

Ethical Risk	Manifestation	Mitigation
Bias	False positives for certain driving styles	Adaptive thresholds, user feedback
Opacity	Driver does not understand alert	Mandatory explanation layer
Over-reliance	Driver defers judgment to system	Reinforcement of advisory role
Loss of agency	Forced restrictions	Explicit override mechanisms

These risks and mitigations will be further validated and prioritised through citizen and stakeholder engagement activities under Action 2.

7.4 Traceability and Logging

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

For this use case, the system maintains structured, non-identifying logs that record trigger conditions in abstracted form, system recommendations, and driver responses such as acceptance or override. These logs are used exclusively for internal quality assurance, ethical performance evaluation under Action 6, and system improvement. No long-term behavioral profiling is performed beyond the context of the individual trip, and no raw biometric data is retained unless explicitly required and consented.

7.5 Relevance to TRUST-AI Objectives

This use case exemplifies how human-in-the-loop governance can be operationalised in practice and how ORRI principles translate into concrete design choices within ADAS systems. It demonstrates why participatory validation is essential for defining acceptable thresholds, explanations, and intervention boundaries. The scenario will be reused as a reference case in citizen foresight workshops under Action 2, in the formulation of the internal ORRI Strategy and Ethics Charter under Action 4, and in the development of evaluation metrics and KPIs under Action 6.

8. Concrete ADAS Use Case Walkthrough (AR-based occlusion-aware situational awareness)

This section presents a second concrete ADAS use case illustrating the application of Responsible AI principles in an Augmented Reality (AR)–based system that enhances driver situational awareness. The use case focuses on visualising otherwise occluded road users or hazards in complex environments, while carefully managing uncertainty, trust calibration, and human control.

8.1 Use Case Description

The use case concerns an AI-powered AR visualisation system that detects the presence of road users or hazards hidden behind physical occlusions, such as buildings, parked vehicles, or large trucks. The system projects this information into the driver's field of view through an in-vehicle AR interface, supporting anticipation rather than automation.

The primary purpose of the system is to reduce collision risk in dense urban and peri-urban environments by making hidden hazards perceptible before they become directly visible. At the same time, the system is explicitly designed to avoid over-reliance, misleading realism, or false authority of visual AI outputs.

The operational context assumes semi-automated driving, with full human responsibility for driving decisions. AR content is provided strictly as decision support and does not trigger automatic vehicle control actions.

8.2 Step-by-Step System Operation

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

During the perception phase (multi-source sensing), the system aggregates information from multiple sensing sources. These include vehicle-mounted sensors such as cameras, LiDAR, and radar, as well as cooperative inputs where available, such as static infrastructure cameras or V2X messages. Vehicle localisation data and digital map context are also integrated. AI models fuse these heterogeneous inputs to infer the probable presence, location, and trajectory of occluded objects, such as pedestrians, cyclists, or other vehicles. A key ORRI consideration at this stage is the explicit computation of uncertainty, which is preserved and propagated through subsequent stages rather than being suppressed.

In the interpretation phase (AI inference and risk estimation), the AI estimates the likelihood that an occluded object is present, predicts its potential motion path, and assigns a confidence score with uncertainty bounds. Only predictions exceeding predefined confidence thresholds are considered for visualisation. The system remains strictly advisory, with no control actions initiated.

In the visualisation phase (AR content generation), the system generates AR content designed to communicate risk without conveying false certainty. Visual elements may include semi-transparent, abstract silhouettes of occluded road users, directional arrows indicating possible motion paths, and colour-coded indicators reflecting confidence levels. Contextual explanatory labels accompany the visualisation, such as "Possible pedestrian detected behind parked vehicle." The AR design deliberately avoids photorealism to prevent the illusion that the system is "seeing" hidden objects with certainty.

Once AR cues are displayed, the driver interprets the information in conjunction with their own perception and driving judgment. The driver decides whether to slow down, change trajectory, or ignore the suggestion. No automatic intervention occurs at this stage. Human agency is preserved by ensuring that AR cues enhance awareness without replacing decision-making.

The system also provides mechanisms for override, calibration, and feedback. The driver can temporarily disable AR overlays, adjust sensitivity settings, or optionally provide feedback on false or unnecessary alerts. System behaviour can adapt within predefined safety and ethical constraints, ensuring responsiveness without uncontrolled learning.

8.3 Ethical Risk Analysis for the Use Case

This use case introduces ethical risks specific to AR-mediated perception. These include over-trust, where drivers may assume AR cues are always correct; opacity, if drivers do not understand why an object appears; misleading realism, if AR visuals appear too concrete; bias, if certain road users are detected less reliably; and cognitive overload from excessive visual elements. Mitigation measures include explicit uncertainty visualisation, contextual explanation labels, abstract non-photorealistic graphics, multi-source sensor fusion, and minimalist, priority-based rendering.

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

Ethical Risk	Manifestation	Mitigation
Over-trust	Driver assumes AR is always correct	Explicit uncertainty visualization
Opacity	Driver does not know why object appears	Contextual explanation labels
Misleading realism	AR appears "too real"	Abstract, non-photorealistic visuals
Bias	Unequal detection of certain road users	Multi-source fusion and validation
Cognitive overload	Excessive AR elements	Minimalist, priority-based rendering

These risks and mitigations will be discussed and validated with citizens and stakeholders in Action 2.

8.4 Data Handling and Governance

For this use case, no personal identification of road users is performed. Predictions are transient, context-bound, and processed in real time. No AR-related data is stored beyond the driving session, unless strictly required for internal system evaluation and explicitly anonymised. All processing follows data minimisation and purpose limitation principles in line with GDPR and ORRI values.

8.5 Traceability and Logging

The system maintains structured logs capturing trigger conditions for AR visualisation, confidence levels and uncertainty ranges, and driver interactions such as enabling or disabling AR overlays. These logs are used exclusively for ethical performance evaluation under Action 6, system improvement, and bias analysis. No behavioural profiling or secondary use of data is performed.

8.6 Relevance to TRUST-AI Objectives

This use case demonstrates how AI-based perception augmentation can be designed to be explainable, uncertainty-aware, and ethically governed. It highlights the novel ethical challenges introduced by AR interfaces, particularly regarding trust calibration and visual authority, and shows why human-in-the-loop design is essential to prevent over-reliance on AI-generated visual cues. The scenario will serve as a reference case for foresight workshops and citizen panels under Action 2, for internal ORRI Strategy development and design rules under Action 4, and for evaluation KPIs related to trust calibration, explainability, and cognitive load under Action 6.

9. Limitations and Next Steps

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 Mapping Report reflects the current design assumptions, system architectures, and development practices of AviSense's ADAS solutions at the outset of the TRUST-AI project. As such, it represents a snapshot of the system as it is currently conceived and implemented, rather than a definitive or final ethical assessment. Certain elements—such as the maturity of specific AI modules, the availability of real-world testing data, and the operational context of deployment—may evolve over the course of the project and influence the relevance or prioritisation of identified risks. A key limitation of this initial mapping is that it is primarily informed by internal technical analysis and design documentation. While this ensures technical accuracy and feasibility, it does not yet incorporate the perspectives, values, and lived experiences of drivers, citizens, or external stakeholders. Addressing this limitation is a central objective of the next project phases.

Accordingly, the Mapping Report will be systematically reviewed and refined through the participatory activities foreseen in **Action 2**, including citizen foresight workshops and the Deliberative Digital Panel. Feedback gathered during these activities will be used to validate identified ethical risk hotspots, challenge underlying assumptions, and highlight additional concerns that may not be evident from a purely technical standpoint. The document will then be updated to reflect stakeholder input, emerging regulatory interpretations, and lessons learned from subsequent actions. In this way, the mapping will function as a **living document** throughout the TRUST-AI project, continuously informing governance design, internal policy development, and evaluation activities.

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