

# Deliverable

## D3.1 Ethical AI Guidelines

WP3 Cross-Domain AI Research Initiative: Advancing Interdisciplinary AI

Title of the work package: Cross-Domain AI

Responsible partner: **DFKI**

Partners: DFKI, GRATEX, IT4I@VSB

[www.innovaite.sk](http://www.innovaite.sk)



Financované  
Európskou úniou  
NextGenerationEU

PLÁN [OBNOVY]



ÚRAD PODPRESEDU VLÁDY  
SLOVENSKEJ REPUBLIKY  
PRE PLÁN OBNOVY  
A ZNALOSTNÚ EKONOMIKU



# Contents

|                                                                                     |    |
|-------------------------------------------------------------------------------------|----|
| 1. Theoretical Framework .....                                                      | 7  |
| 1.1 Rationale .....                                                                 | 7  |
| 1.2 Theoretical Framework Based on the HLEG Principles .....                        | 9  |
| 2. HLEG and ALTAI: Bridging Normative Principles and Practical Implementation ..... | 11 |
| 3. InnovAlte Ethics Evaluation: Pre-Process .....                                   | 12 |
| 4. InnovAlte Ethics Evaluation Framework .....                                      | 14 |
| 5. InnovAlte Incident Knowledge Base .....                                          | 17 |
| Ongoing Ethics Evaluation and Follow-Up .....                                       | 19 |
| ANNEXES .....                                                                       | 20 |
| Annex 1 - WP4 .....                                                                 | 21 |
| Annex 2 - WP5 .....                                                                 | 25 |
| Annex 3 - WP6 .....                                                                 | 33 |
| Annex 4 - WP7 .....                                                                 | 36 |

| Revision history |            |         |                        |
|------------------|------------|---------|------------------------|
| Version          | Date       | Comment | Author                 |
| v1               | 31.07.2026 |         | DFKI                   |
| v2               |            |         | DFKI, GRATEX, IT4I@VSB |
|                  |            |         |                        |

# Executive Summary

InnovAlte Slovakia develops AI solutions across several domains:

## 1. WP4: Sustainable Energy Strategy (Dynamic Efficiency Control)

WP4 focuses on AI-driven energy management. Its objective is to develop smart control systems for buildings and urban energy networks that optimize consumption. The work covers sensing, forecasting, and control algorithms to adapt heating, cooling, and other energy use in real time. Expected outcomes include demonstrator applications that significantly reduce energy waste and environmental impact. For example, WP4 aims to deliver an AI-based controller that learns from user patterns and weather forecasts to minimize heating/cooling needs. All technological components developed within this work package target a maximum Technology Readiness Level (TRL) of 5.

## 2. WP5: AI-Video Safety Systems

WP5 addresses safety and comfort in buildings and roads through AI-based video analytics. Its objectives are: (1) develop AI vision systems to monitor building infrastructure (occupancy, environment conditions, and safety and security hazards) to enhance comfort, protection and energy use; and (2) use AI-driven video to detect road or traffic hazards for improved traffic safety. Expected outcomes include demonstrations of novel AI surveillance applications that can, for instance, alert to security breaches, safety threats, or unsafe conditions, as well as optimize building operations for user comfort. All technological components developed within this work package target a maximum Technology Readiness Level (TRL) of 5.

## 3. WP6: Automating Software Development with AI

WP6 focuses on the automation of the software engineering process using AI. Its objectives include creating tools and methods that allow AI to participate in coding tasks. For example, auto-generating code, assisting debugging, or suggesting optimizations. The scope covers the entire development lifecycle, aiming to enhance efficiency and quality of software products. Expected outcomes are software prototypes or modules that significantly speed up development and reduce errors (e.g. AI-driven code generators or intelligent IDE assistants). All technological components developed within this work package target a maximum Technology Readiness Level (TRL) of 5.

## 4. WP7: AI-Driven Insurance Operations

WP7 aims to embed AI into insurance back-end operations. Its objectives are to optimize claims handling, underwriting, and customer-service processes using AI (such as machine

learning for risk assessment or natural language processing for claim documents). The work package's scope includes developing AI modules for data processing, fraud detection, and automated decision support. Expected outcomes are demonstrator systems that speed up insurance workflows and improve accuracy. All technological components developed within this work package target a maximum Technology Readiness Level (TRL) of 5.

## 5. Planned WP8: AI-Infused Prototypes and Tangible Advancements

WP8 is a planned integration and prototype development package. Its purpose is to combine the outputs of WPs 4-7 into prototypes ready for market demonstration. Activities under WP8 include system integration, testing, and validation of cross-domain AI systems. WP8 will handle how the various AI systems (energy management, safety monitoring, development tools, insurance platforms) can be packaged into prototypes. These prototypes (sometimes called "infused prototypes") will be demonstrations of the technology. For example, an energy control system from WP4 could be integrated into a smart-building demo, or an AI claims automation pipeline from WP7 could be tested with industry partners. Thus, WP8's outcome will be cross-sector prototypes and modules that industries can readily deploy or pilot. WP8's foundational objective is the maturation of AI technologies to attain Technology Readiness Levels (TRL) 6-7.

Because all the above-mentioned systems differ in purpose, data, and risk profile, the project developed and uses an ethics framework based on the HLEG Ethics Guidelines for Trustworthy AI and ALTAI. This approach ensures that ethical requirements are applied consistently and at the same time are still being adapted to each system's specific context.

WP3 established a three-step ethics evaluation process: system characterisation, preliminary risk mapping, and review with recommendations. In addition, the document contains an Incident Knowledge Base to support early risk awareness and responsible development. Together, these measures help ensure that InnovAite's AI systems are developed in a trustworthy, transparent, and accountable way.



# Introduction

The rapid expansion of AI technologies across economic and social domains has increased the need for robust ethical governance frameworks that can guide not only the design of AI systems, but also their development, testing, deployment, and ongoing reassessment. In interdisciplinary projects such as InnovAite, the challenge is not the absence of ethical principles, but the diversity of application contexts and the risk that ethical requirements remain fragmented, abstract, or inconsistently applied. A coherent framework is therefore needed to translate normative commitments into operational safeguards that are meaningful across very different technical systems.

This deliverable responds to that need by defining the ethical framework for InnovAite and by setting out the methodology through which ethical review will be carried out across the consortium. The project adopts the HLEG Ethics Guidelines for Trustworthy AI as its normative foundation because they are widely recognised in the European policy landscape, aligned with emerging regulatory expectations, and accompanied by practical implementation tools. The accompanying ALTAI checklist is used as the main assessment instrument, but not as a rigid one-size-fits-all template. Instead, the checklist is adapted to the characteristics and risks of each system so that the evaluation remains both rigorous and context sensitive.

The methodological core of the deliverable is the InnovAite Ethics Evaluation Framework developed in WP3. It begins with system characterisation, in which each work package provides structured information on system capabilities, intended use, data dependencies, development stage, deployment setting, and human oversight. The annexes demonstrate how the framework is applied to concrete project research.

All InnovAite WPs in which AI constitutes a core technological component have been included for the ethical evaluation process. These work packages undergo the evaluation using the InnovAite Ethics Evaluation Framework and the adapted ALTAI checklist. Work packages that do not develop or deploy AI systems within the meaning of the AI Act are not subject to a full AI ethics evaluation, although they continue to comply with the project's broader ethical, legal, and data governance requirements where applicable.

# 1 Theoretical Framework

## 1.1 Rationale

The rapid integration of AI in different societal areas has intensified the need for robust ethical governance frameworks capable of guiding the design, development, and deployment of AI systems. While technological capabilities have expanded rapidly, the societal implications of AI (ranging from the risk of algorithmic bias and privacy violations to concerns regarding transparency, accountability, and human autonomy) have prompted policymakers, researchers, and industry stakeholders to seek systematic approaches to responsible innovation. In this context, the development of a coherent ethical framework is not merely a matter of compliance with policy obligations, but an essential component of trustworthy technological progress. AI systems increasingly shape decisions that affect individuals, organisations, and entire socio-economic systems; therefore, their development must be guided by clearly articulated ethical principles that ensure alignment with fundamental rights, societal values, and democratic norms.

The challenge faced by many contemporary research initiatives lies not in the absence of ethical principles but in the fragmentation of the normative landscape. Numerous frameworks, guidelines, and sector-specific codes exist, often overlapping in their terminology and objectives but differing in their structure, scope, and operationalisation. When multiple principles are assembled without a single normative anchor, ethical governance risks becoming conceptually inconsistent and difficult to implement. Such fragmentation complicates the process of translating abstract values into verifiable technical and organisational requirements, and it weakens the ability of projects to demonstrate coherent ethical reasoning to external evaluators, regulators, and the wider public. For complex interdisciplinary projects involving multiple partners and diverse application domains, the absence of a unified ethical framework can therefore lead to ambiguity, uneven implementation of safeguards, and limited transparency regarding the ethical rationale underlying technological decisions.

To address this challenge, the project adopts as its ethical foundation the principles articulated by the High-Level Expert Group on Artificial Intelligence in the document Ethics Guidelines for Trustworthy AI. This framework represents the most widely recognised ethical reference for human-centric AI within the European policy landscape. Developed through an extensive multi-stakeholder process, the guidelines articulate a comprehensive vision of trustworthy artificial intelligence that integrates ethical reflection with practical governance mechanisms. The framework identifies seven core requirements for trustworthy AI: human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; societal and environmental well-being; and accountability. Together, these principles establish a normative architecture that connects technological development with the protection of fundamental rights and the promotion of socially beneficial innovation.

The choice to adopt this framework is motivated by several considerations that extend beyond its conceptual clarity. First, the HLEG guidelines possess strong normative legitimacy within the European research and policy environment. They are widely recognised as the central ethical reference point for responsible AI development and have influenced numerous national and international governance initiatives. Second, the framework has direct regulatory relevance, particularly in light of the evolving European legal architecture for artificial intelligence. The EU Artificial Intelligence Act explicitly references the work of the HLEG and incorporates governance concepts that closely reflect the principles articulated in the guidelines, including requirements related to transparency, risk management, and fundamental rights assessments for high-risk AI systems.

A further advantage of the HLEG framework lies in its emphasis on operationalisation. Ethical principles in research contexts often remain abstract, limiting their practical impact on technological design. In contrast, the HLEG guidelines are accompanied by concrete implementation tools that enable the systematic assessment of AI systems throughout their lifecycle. In particular, the Assessment List for Trustworthy AI (ALTAI) provides a structured checklist that translates ethical principles into evaluative questions and implementation criteria. Such tools make it possible to transform normative commitments into verifiable procedures, thereby enabling to integrate ethical reflection directly into development processes, risk assessments, and validation activities.

Finally, grounding the project's ethical governance in the HLEG framework significantly enhances the transparency and auditability of research practices. Structuring ethical decision-making around a recognised set of principles and documented assessment procedures, the project generates traceable artefacts (such as risk assessments, documentation and evaluation reports). This level of documentation strengthens the demonstrability of ethical compliance and contributes to the reproducibility and credibility of research outcomes. In large collaborative projects, where multiple institutions contribute to the development of complex AI technologies, such transparency is essential for ensuring consistent ethical standards and maintaining public trust.

Against this background, the following sections establish the theoretical ethical framework guiding the project. Building upon the normative structure provided by the HLEG principles, the framework articulates how ethical requirements can be systematically integrated into AI research and innovation processes. It examines the conceptual foundations of trustworthy AI, outlines mechanisms for translating ethical principles into operational safeguards, and proposes governance structures that support accountability, transparency, and responsible technological development. Through this approach, the project seeks to ensure that its scientific and technological advances are pursued in a manner consistent with European ethical values, regulatory expectations, and the broader goal of responsible innovation.

## 1.2 Theoretical Framework Based on the HLEG Principles

Building upon the rationale presented above, the project adopts the ethical architecture articulated by the High-Level Expert Group on Artificial Intelligence in the Ethics Guidelines for Trustworthy AI as the conceptual and operational foundation for responsible AI development. The framework proposed by this expert group defines trustworthy artificial intelligence as a system that simultaneously satisfies three cumulative conditions throughout its lifecycle: it must be lawful, complying with all applicable regulations and legal norms; ethical, respecting fundamental rights and societal values; and robust, both from a technical and social perspective, ensuring reliability and resilience in real-world contexts. These three dimensions form an integrated model of governance in which ethical reflection should be embedded directly within the technological design process rather than being treated as an external or post hoc evaluation.

Within this conceptual structure, the HLEG framework identifies seven key requirements that operationalize the concept of trustworthy AI. These requirements function as normative pillars guiding both the development and deployment of AI systems and provide a systematic approach for identifying and mitigating potential ethical risks. Each requirement addresses a specific dimension of ethical governance while remaining interconnected with the others, forming a comprehensive framework for responsible AI innovation.

The first principle, **human agency and oversight**, emphasizes that AI systems must support rather than undermine human autonomy. AI should enhance human decision-making capacities while preserving meaningful human control over critical processes and outcomes. In practice, this implies the design of systems that allow human intervention, supervision, and the possibility to override automated decisions where necessary. Maintaining human oversight is particularly important in contexts where AI influences safety-critical operations or decisions with significant social consequences. Through embedding mechanisms such as human-in-the-loop, human-on-the-loop, or human-in-command architectures, AI systems can remain aligned with human values and institutional accountability structures.

The second requirement, **technical robustness and safety**, addresses the need for AI systems to function reliably and securely throughout their operational lifecycle. AI technologies must be resilient to errors, adversarial manipulation, and unexpected environmental conditions. Robustness therefore includes not only the technical accuracy of models but also their stability, fault tolerance, and capacity to handle uncertainty. From an ethical perspective, technical robustness is closely linked to the prevention of harm: systems that fail unpredictably or behave inconsistently can undermine public trust and generate unintended societal consequences. Consequently, robust testing procedures, risk management strategies, and continuous monitoring mechanisms must accompany the deployment of AI technologies.

The third principle, **privacy and data governance**, reflects the importance of protecting individuals' personal data and ensuring responsible data management practices. AI systems frequently rely on large datasets that may include sensitive or personally identifiable information.

The ethical governance of such data requires strict adherence to established data protection principles such as data minimization, purpose limitation, and security safeguards. At the European level, these principles are reinforced by the General Data Protection Regulation, which establishes legal requirements for the lawful processing of personal data. Within the HLEG framework, privacy protection is complemented by broader data governance considerations, including data quality, integrity, provenance tracking, and transparent documentation of data collection and processing methods.

A fourth key requirement is **transparency**, which aims to ensure that the functioning and implications of AI systems remain understandable to relevant stakeholders. Transparency encompasses several dimensions, including traceability of datasets and algorithms, explainability of automated decisions, and the clear communication of system capabilities and limitations. From an ethical standpoint, transparency is necessary for enabling accountability and fostering trust between technology developers, users, and affected individuals. Although full technical explainability may not always be achievable for complex machine learning systems, appropriate levels of interpretability and documentation must be provided to ensure that stakeholders can meaningfully assess the behaviour and implications of AI systems.

The fifth requirement, **diversity, non-discrimination, and fairness**, addresses the risk that AI systems may reproduce or amplify existing social inequalities. Machine learning models trained on biased or unrepresentative data can produce discriminatory outcomes that disproportionately affect certain demographic groups. Ethical AI governance therefore requires proactive efforts to identify and mitigate bias during the design, training, and evaluation phases of AI systems. This includes ensuring diversity in training datasets, monitoring system performance across different population groups, and implementing fairness-aware evaluation metrics. Beyond technical considerations, this principle also highlights the importance of inclusive design processes that involve diverse perspectives in the development and governance of AI technologies.

The sixth principle, **societal and environmental well-being**, broadens the ethical perspective beyond immediate technological performance to consider the broader societal impact of AI innovation. AI systems should contribute positively to societal goals such as sustainability, social inclusion, and economic development while avoiding negative externalities that may harm individuals, communities, or the environment. This requirement encourages developers to consider the long-term implications of AI technologies, including their environmental footprint, potential labour market impacts, and broader societal consequences. In this sense, ethical AI development is closely connected to the concept of responsible innovation, which emphasises anticipation, reflection, and responsiveness in technological development.

Finally, the seventh requirement, **accountability**, establishes the governance mechanisms necessary to ensure that responsibility for AI systems can be clearly assigned and enforced. Accountability requires the existence of clear processes for auditing AI systems, documenting design decisions, and addressing potential harms or unintended outcomes. Developers and

organisations deploying AI technologies must be able to demonstrate compliance with ethical standards and provide mechanisms for redress when negative impacts occur. In practical terms, this may involve maintaining comprehensive documentation of development processes, conducting regular risk assessments, and establishing oversight structures capable of monitoring the ethical performance of AI systems over time.

## 2. HLEG and ALTAI: Bridging Normative Principles and Practical Implementation

To ensure that the ethical principles outlined above are translated into concrete and verifiable practices, the project will rely on the structured assessment methodology provided by the Assessment List for Trustworthy AI (ALTAI). Developed as a practical implementation tool accompanying the Ethics Guidelines for Trustworthy AI of the High-Level Expert Group on Artificial Intelligence, ALTAI provides a comprehensive self-assessment framework designed to operationalise the seven requirements of trustworthy AI. Rather than functioning as a purely theoretical instrument, ALTAI translates each ethical principle into a structured set of evaluative questions, indicators, and documentation requirements. This approach enables to assess whether AI systems meet the ethical standards established within the HLEG framework and to identify potential risks or gaps that may require mitigation.

Furthermore, the adoption of ALTAI contributes to the alignment of the project with the broader European regulatory environment. As it is the case with HLEG 7 Principles, the assessment logic underlying ALTAI reflects governance principles that are increasingly embedded within European AI regulation, particularly in the EU Artificial Intelligence Act. Through employing this methodology, the project anticipates emerging compliance expectations related to risk assessment, transparency obligations, and the evaluation of potential impacts on fundamental rights. Consequently, ALTAI serves not only as an ethical governance instrument but also as a strategic tool supporting regulatory readiness and responsible innovation.

Through the systematic use of ALTAI, the project operationalises the theoretical ethical framework established by the HLEG principles. The checklist functions as a bridge between normative commitments and practical implementation, ensuring that ethical values such as fairness, accountability, transparency, and human oversight are embedded within the technical, organisational, and governance processes of the project. In this way, ethical evaluation becomes an ongoing and structured component of the research lifecycle, reinforcing the project's commitment to the development of trustworthy, human-centric AI.

As the project's subchapters below will demonstrate, the diversity of systems and technologies involved in InnovAite project makes it necessary to adapt the ALTAI checklist for each case in

part. ALTAI provides a strong, structured foundation for evaluating project research and results which will be systematically curated for each specific system or technology developed. Curations will adapt question-priorities, add domain-relevant checks, and set concrete acceptance thresholds that reflect the particular harms. These tailored lists do not replace ALTAI, the original ALTAI checklist remains the baseline grammar and normative anchor, while the curated variants furnish the domain-sensitive dialects necessary for meaningful evaluation.

### 3. InnovAlte Ethics Evaluation: Pre-Process

The development of the Ethics Evaluation Process within WP3 began with an analysis conducted by the DFKI Ethics Team: a structured questionnaire was circulated to all work packages to collect concise system dossiers. It comprised 12 questions that sequentially map the lifecycle and sociotechnical footprint of each AI system, from domain scope and concrete use cases (sections 1-3), through technical architectures and functional features (sections 4-6), to expected outputs, data flows, stakeholder roles, data-sharing practices, risk profiles and training needs (sections 7-12).

The exercise served three purposes: to map the research domains of InnovAlte and the differing constraints across domains, to surface the recurring ethical challenges, and to identify which evaluative instruments would be both practicable and informative for each WP.

The analysis of the questionnaire responses generated two results: first, that project's modalities range widely (from building control loops and synthetic-data rendering to LLM-assisted software tools and regulated insurance systems), and second, that a single unadapted (ALTAI, or other) checklist would either miss domain-specific risks or impose irrelevant requirements. These findings directly motivated the decision to use a contextualized assessment for each developed system.

Against that backdrop, the project adopted the ALTAI instrument as the foundational scaffolding for evaluation: ALTAI's strength lies in its ability to translate high-level HLEG requirements into disciplined, evidence-seeking questions that demand documented artefacts rather than assertions. However, the questionnaire's findings also clarified that ALTAI must be treated as an adaptable baseline rather than a one-size-fits-all instrument.

Therefore, the DFKI Ethics Team will curate ALTAI-derived lists for each system or technology developed within InnovAlte: starting from the core ALTAI questions, the Ethics Team performs a targeted domain mapping (identifying domain-specific harm vectors, applicable constraints, and realistic mitigation options), adjusts the emphasis of ALTAI items, and delivers to InnovAlte research groups specific ethics evaluation lists for their system/technology. This curated approach preserves ALTAI's methodological rigor but is also ensuring that evaluations are sensitive to the concrete, heterogeneous risks that the InnovAlte work packages present.



## 4. InnovAlte Ethics Evaluation Framework

The InnovAlte Ethics Evaluation Framework establishes the methodological structure through which ethical aspects of AI systems developed within the project are assessed. Given the technological diversity of the project, ethical evaluation cannot rely on ad hoc assessments or isolated compliance checks. Instead, the project requires a structured and reproducible approach capable of identifying risks, verifying safeguards, and ensuring that AI development remains aligned with a responsible innovation approach.

WP3 developed a framework for ethics evaluation that responds directly to the project objectives and is built from three inputs: the theoretical framework of InnovAlte (the HLEG principles), empirical evidence gathered through the project questionnaire, and the practical assessment tooling adapted for project needs. The DFKI Ethics Team will mediate the translation from principle to practice, producing a reproducible evaluation structure that researchers can engage with as part of regular development cycles. The evaluation framework is conceived to integrate research design, data practices, and deployment decisions into a single evaluative loop.

As illustrated in Figure 1, the InnovAlte Ethics Evaluation Framework consists of three sequential stages: (1) system characterization, (2) ethical risk screening and selection of relevant ALTAI questions, and (3) evaluation, recommendations, and iterative follow-up.

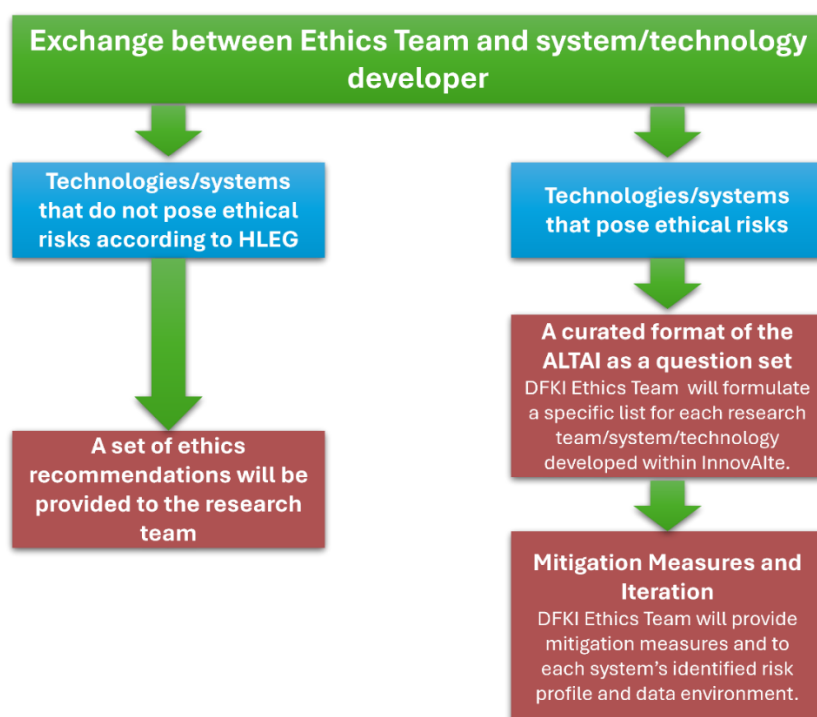


Figure 1. InnovAlte Ethics Evaluation Framework illustrating the three-stage ethics evaluation process

**The first stage of the evaluation** focuses on system characterization and contextual understanding. At this stage, each work package provides a structured description of the systems or technologies under development using a questionnaire. The purpose of this step is to establish a clear picture of the system's capabilities, intended use, data dependencies, and operational context before any ethical evaluation is performed. Without this initial mapping, the identification of risks and appropriate mitigation strategies would lack the necessary technical and contextual grounding.

The description begins with an analysis of the system's capabilities. Work packages are asked to specify the type of system being developed (for example, whether the technology functions as a decision-support tool, an automation system, a predictive model, a data analytics platform, or a generative AI application). This distinction is important because different system architectures present different ethical and governance challenges. In addition to identifying the system category, the questionnaire requires clarification of the degree of autonomy present in the system. In particular, developers must indicate whether the system performs certain tasks or decisions fully autonomously or whether automated outputs are designed to support human decision-making. Understanding the extent of automation allows the DFKI Ethics Team to determine the potential implications for human oversight and accountability.

The second component of the description concerns the intended use of the system. Here, work packages explain the scenarios in which the technology is expected to operate. These descriptions typically include operational environments, user groups, and the specific problems the system is designed to address. Documenting the intended use is needed because ethical risks are often context-dependent: a technology that is harmless in a research environment may have significant consequences when deployed in operational settings.

The third component addresses data sources and data characteristics. Work packages must describe the types of data used by the system, including whether the data are synthetic, historical, derived from operational environments, etc. The questionnaire also asks developers to identify the origins of these datasets and the categories of information they contain. This information is needed for assessing privacy implications and data governance requirements, particularly when personal or sensitive data may be involved.

Finally, the system description addresses the deployment context and governance structure. Work packages must indicate the current stage of development (for example, research prototype, experimental validation, or pilot testing) and describe the environment in which the system will be deployed. This includes whether the system operates within controlled experimental settings, internal organizational environments, or pilot deployments with external stakeholders. Equally important is the clarification of human oversight mechanisms. Developers must specify whether the system operates under a human-in-the-loop, human-on-the-loop, or supervisory monitoring model and identify the actors responsible for reviewing outputs and making final decisions. Establishing these oversight mechanisms at the outset allows the ethics

evaluation to assess accountability structures and determine whether sufficient human control is maintained over automated processes.

**The second stage of the evaluation** focuses on the identification and preliminary mapping of ethical risks associated with the system described in the first stage. Once the answers to the questions from the first phase have been received, the Ethics Team analyses the information to determine which ethical dimensions are relevant for the technology under consideration. This stage functions as a structured screening phase that identifies the primary areas of concern and determines the level of scrutiny required for the subsequent evaluation.

During this screening, the Ethics Team examines the system through the HLEG dimensions. These may include potential privacy and data protection issues, risks of bias and discrimination, accuracy and reliability limitations, the possibility of over-reliance on automated outputs, and broader concerns related to transparency, explainability, and accountability. Particular attention is given to the role that automated outputs may play in decision-making processes and whether these outputs could have direct or indirect consequences for individuals or organizations. The analysis also evaluates whether the system processes personal or sensitive data, whether it interacts with human users in operational contexts, and whether its outputs could influence safety-critical or economically significant processes.

At this stage, the evaluation also considers the technical characteristics of the system and the data ecosystem in which it operates. For example, systems relying on generative models or large language models may raise concerns related to hallucinations, unreliable outputs, or misuse through prompt manipulation. Systems that process historical datasets may present risks related to inherited bias or insufficient anonymization, while operational systems may introduce risks associated with automation errors, cybersecurity vulnerabilities, or incorrect recommendations that affect real-world processes.

This second stage of the evaluation will split the evaluated systems in two categories:

1. Technologies/systems that do not pose ethical risks, which together with the text of the ethics evaluation receive a set of optional recommendations.
2. Technologies/systems that pose ethical risks, which together with the text of the ethics evaluation receive A curated format of the ALTAI as a question set to be completed.

For the second category, based on the identified risks and system characteristics, the Ethics Team compiles a curated list of relevant evaluation questions derived from the ALTAI List developed by the HLEG. Rather than applying the entire ALTAI checklist uniformly, the team selects those questions that are applicable to the specific system under evaluation and, where necessary, adapts them to reflect the technical and operational context of the project. This curated questionnaire ensures that the assessment remains focused on the most relevant ethical dimensions while avoiding unnecessary evaluation steps that do not apply to a particular technology.

The **third stage of the evaluation** operationalises the findings from the curated ALTAI assessment and provides recommendations and, if needed, mitigations measures. After the DFKI Ethics Team has sent the curated ALTAI questionnaire to the corresponding WP and reviewed the WP's responses, the Ethics Team conducts an evaluation of those answers. The Ethics Team documents the outcomes of the review together with recommendations for the continued development of the system.

The recommendations are intended to assist researchers in maintaining good practices, strengthening existing safeguards, and anticipating ethical considerations that may become relevant as the system matures. The recommendations also provide a basis for future reassessment should the technical architecture, data sources, operational environment, or intended use of the system evolve.

This process is collaborative and iterative. The DFKI Ethics Team encourages researchers to maintain regular communication between researchers and ethics advisors, as AI systems are often subject to continuous refinement. Significant changes to datasets, functionalities, deployment assumptions, stakeholder groups, or research objectives should trigger renewed consultation with the Ethics Team. Through this ongoing dialogue, the project can ensure that ethical oversight remains proportionate to the evolving characteristics of the system and that the documented ethics response continues to reflect the realities of the research and development process.

The entire lifecycle for each assessed system is therefore reconstructible: initial screening, curated ALTAI responses, and results. This provenance trail supports external reviewers in reconstructing ethical decision-making pathways. The output is captured both operationally and as deliverable material: the Ethics Team produces a formal dossier for each evaluated system. Questionnaires and reports will be attached as annexes to this deliverable (living document).

## 5. InnovAite Incident Knowledge Base

The InnovAite Incident Knowledge Base is formulated as a proactive repository that the consortium will consult at the outset of any design or development activity and during iterative reassessments. Its purpose is to make prior lessons, empirical evidence, guidance, and tested mitigation patterns plainly available so teams can anticipate hazards and embed protections before code is written or datasets gathered. Rather than functioning as a static bibliography, the Knowledge Base is organised as an actionable toolset.

The Knowledge Base will, by design, consist of a curated catalogue of authoritative public online AI incident databases rather than an internal archive of raw incident reports; in other words, the repository that project teams can consult proactively before starting development will be a maintained list of external, community-trusted feeds and trackers together with short guidance on their scope, reliability, and best use cases. This approach leverages existing collective data

about real-world AI failures and near-misses, avoids unnecessary duplication of effort, and ensures that work packages begin design and data collection with current empirical knowledge about failure modes and the mitigations that have been shown to work.

The primary external resources to be listed and routinely monitored are summarized in Table 1:

|   |                                                                                |                                                   |                                                                                                                                                                   |                                                                    |                                                              |                                                                       |
|---|--------------------------------------------------------------------------------|---------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------|--------------------------------------------------------------|-----------------------------------------------------------------------|
| 1 | AI Incident Database (AIID)                                                    | Partnership on AI community initiative            | Real-world AI failures and harms: deepfakes, algorithmic bias, autonomous vehicle incidents, chatbot misuse, fraud, misinformation                                | ~2,800+ incident reports (continuously growing)                    | Public incident registry with structured taxonomy            | <a href="https://incidentdatabase.ai">https://incidentdatabase.ai</a> |
| 2 | AIAAIC (AI, Algorithmic and Automation Incidents and Controversies Repository) | Responsible AI Collaborative                      | Incidents and controversies involving AI, algorithms, and automated decision systems: discrimination, privacy violations, misinformation, and governance failures | ~1,000+ documented cases                                           | Incident and controversy repository with categorized entries | <a href="https://www.aiaaic.org">https://www.aiaaic.org</a>           |
| 3 | MIT AI Incident Tracker                                                        | Massachusetts Institute of Technology AI Risk Lab | Analytical tracking of AI incidents by sector and harm type (privacy, safety, bias, misinformation)                                                               | Aggregated analysis of hundreds of incidents from multiple sources | Analytical dashboard and trend analysis tool                 | <a href="https://airisk.mit.edu">https://airisk.mit.edu</a>           |
| 5 | Algorithm Watch Case Studies and Datasets                                      | Algorithm Watch                                   | Algorithmic accountability investigations: bias in automated decision systems, recommendation systems, predictive policing, and platform algorithms               | Dozens of curated investigations and datasets                      | Research reports and case-study datasets                     | <a href="https://algorithmwatch.org">https://algorithmwatch.org</a>   |

**Table 1. Curated public AI incident repositories included in the InnovAite Incident Knowledge Base.**

## Ongoing Ethics Evaluation and Follow-Up

The InnovAlte Ethics Evaluation Framework is designed as a living process rather than a one-time assessment. As the AI systems developed within the project mature, their architectures, functionalities, data sources, and intended uses may evolve. Consequently, the ethical considerations, identified risks, and corresponding mitigation measures described in this deliverable should also be regarded as dynamic and subject to continuous review. The present document therefore constitutes a living document that will be regularly updated to reflect the evolution of the project, the outcomes of subsequent ethics evaluations, and newly identified risks or mitigation measures.

Ethical oversight within InnovAlte will therefore continue throughout the research and development lifecycle and will ensure that ethical safeguards remain proportionate to the evolving characteristics of the systems and continue to reflect the principles of trustworthy AI established by the HLEG Ethics Guidelines and operationalized through the adapted ALTAI methodology.

## ANNEXES

### Explanatory note on the Annexes

Each work package is accompanied by a separate annex that documents the application of all three steps of the InnovAite Ethics Evaluation Framework to the specific system or technology developed within that WP. All annexes follow the same methodological structure, so that the assessment can be compared across WPs.

In line with the framework set out in the main body of this deliverable, each annex is organised into three corresponding steps:

**Step 1** presents the questionnaire filled by each WP on the system characterisation and contextual understanding. The questionnaire provides the description of each system capabilities, intended use, data sources, deployment context, and human oversight arrangements. The answers of the respective WPs are written in italic text.

**Step 2** describes the preliminary identification and mapping done by the DFKI Ethics Team on the ethical risks of the respective systems/technologies. For WP4 and WP6, the preliminary identification includes a set of recommendations based on the identified risk profile. For WP5 and WP7, (due to the system complexity) the preliminary identification is continued and extended in the same document with a curated list of questions derived from the ALTAI List, designed to support a more detailed and context-specific assessment of the relevant ethical dimensions. The answers of the respective WPs to the curated list of questions are also attached in italic text.

**Step 3** covers the ethics review outcomes and (if applicable) recommendations, following the exchange had between the Ethics Team and the respective WPs in Step 2. The document consolidates the results of the exchange had during the evaluation done during the 2<sup>nd</sup> Step of the Ethics Evaluation Framework.

## Annex 1 - WP4

### WP4 - STEP 1: System Characterisation and Contextual Understanding

#### 1. Capabilities

##### a. What type of system will you develop?

**Answer:**

- *Final output: Data Analytics Platform (software + methodology) designed to detect and quantify inefficient energy consumption while pinpointing the exact time and location of occurrence. Users integrate meter readings or reference production data to prioritize energy-saving measures and assess the return on investment (ROI) for specific interventions.*
- *Internal modules:*
  - *Automated Data Preparation: Handles data conversion and basic/advanced preprocessing. Uses correlation analysis to select relevant datasets.*
  - *Digital/Predictive Model: Generates consumption models via trained digital twins or standard/average reference values.*
  - *Inefficiency Calculation: Processes real vs. expected consumption to identify various types of energy waste.*
  - *Visualization: Combines standard graphical interfaces with BIM (Building Information Modelling) integration.*
- *Optional modules:*
  - *What-if simulation for basic scenarios (e.g., adding PV/BESS).*
  - *Automated configuration optimization.*
  - *Predictive control simulation – generates control strategies and calculates potential scenario outcomes.*

##### b. In which areas does the system make decisions entirely autonomously?

**Answer:** *In its base configuration, the system does not make any autonomous control decisions. Its purpose is to pre-process information so the user can quickly and continuously evaluate energy efficiency. To reduce operational complexity, some data adjustments are automated; however, the user is prompted for input/configuration whenever the data strategy is unclear.*

*Note: We are considering a simulation predictive control module to generate instructions, but direct connection to control infrastructure is not planned.*

## 2. Intended use

### a. In what scenarios will the system be used?

**Answer:**

- *When a user needs to determine if their energy efficiency is significant enough to warrant corrective measures.*
- *When a user needs to prioritize energy optimization tasks and determine financial payback periods.*

## 3. Data sources

### a. What types and sources of data does the system use?

**Answer:**

1. *Energy consumption data: Meter readings and measurements.*
2. *Reference power data: Production volumes, process inputs, or other contextual data.*
3. *External data: Spot market prices, weather forecasts, and centralized energy metrics (e.g., CO2 intensity).*

## 4. Deployment context

### a. What is the current development stage?

**Answer:** Individual sub-components are currently being developed by various WP4 teams. Simultaneously, a shared coordination software platform is under development to integrate these outputs using a microservices architecture.

- **Deployment context:** The project collaborates with 5 reference users, and we are currently processing the initial datasets.

### b. What level of human oversight exists, and who is accountable for outcomes?

**Answer:**

- *Level of Oversight: Human-in-the-Loop (HITL). The human is a direct part of the process. The system acts as a decision-support tool and cannot execute any primary actions or final reports without explicit human approval.*
- *Accountability: The end-user is ultimately accountable. The system serves strictly as an analytical support framework for the user's decision-making process.*

## WP4 - STEP 2: Preliminary identification and mapping of ethical risks

Based on the current system description, the energy data analytics platform does not raise significant risks considering ALTAI dimensions. In particular, the system is not designed to process personal data or special-category data, it does not make autonomous decisions, and it does not connect to operational control infrastructure.

### 1. Privacy and data governance

The system does not pose a privacy risk because it is designed to process energy consumption data, production-related reference data, and external contextual data. Therefore, it does not appear to create a direct risk of unlawful processing of personal or special-category data, provided that implementation remains limited to the described data scope.

### 2. Technical robustness and safety

The platform does not pose a direct safety risk in the sense of autonomous physical action, because it does not control equipment or execute operational commands. Its role is limited to preprocessing, analysis, visualisation, and support for human decision-making. For this reason, a failure in the system would not directly trigger physical harm through automated control. The remaining risk is analytical rather than safety-critical: the system could still produce incorrect estimates if the data or model assumptions are flawed. However, this is a risk of result quality, not of direct autonomous harm. The system therefore does not introduce the type of technical safety risk associated with closed-loop control systems.

### 3. Human agency and oversight

The system is explicitly designed as a decision-support tool and not as an autonomous decision-maker. Final judgments remain with the user, and the platform does not make primary decisions without human approval. This means that the system does not undermine human agency in the way that fully automated systems might. The human-in-the-loop structure must ensure that users retain control over interpretation, prioritisation, and implementation of energy-saving measures.

### 4. Transparency

The platform is intended to explain energy inefficiency through consumption comparisons, reference values, and visualisation tools, which are inherently documentable and reviewable. It is not a system that generates opaque decisions about people, access, eligibility, or classification. The methodological transparency is still required, but the system does not create the specific transparency risk associated with hidden decision logic in high-impact AI systems.

## 5. Accountability

The system does not pose a major accountability risk caused by unclear responsibility for autonomous decisions, because it does not autonomously act on the world or issue binding final decisions. Responsibility remains with the user and the organisation operating the tool. The platform supports analysis and prioritisation, but it does not itself approve investments, change operational settings, or enforce actions. As a result, the accountability chain is straightforward: the system provides support, while humans remain accountable for interpretation and action. The platform therefore does not create the kind of accountability gap associated with autonomous AI systems.

### Recommendations:

As a general recommendation, the system should remain within a clearly defined decision-support scope, with documented data provenance, transparent calculation logic, and routine human review of outputs before any action is taken. Basic validation of input data, model assumptions, and scenario results should be maintained throughout development and deployment to ensure that the platform remains reliable and interpretable.

## Annex 2 – WP5

### WP5 - STEP 1: System Characterisation and Contextual Understanding

#### 1. Capabilities

##### a. What type of system will you develop?

**Answer:** *AI systems in the form of predictive models (inference components) for applications in smart buildings and traffic safety. These models focus on computer vision tasks such as object detection, human activity recognition, and anomaly detection from visual data.*

##### b. In which areas does the system make decisions entirely autonomously?

**Answer:** *None. The systems are explicitly limited to the prediction/inference layer and do not perform autonomous decision-making, actuation, or control. They produce probabilistic outputs (e.g., classifications, detections, and anomaly scores) from input data.*

#### 2. Intended use

##### a. In what scenarios will the system be used?

**Answer:** *The systems will be used for visual data analysis in contexts such as smart buildings and traffic environments. Their functions are limited to generating predictive outputs (e.g., detection and classification results) that may be used by external systems or human operators. The systems themselves do not provide recommendations or trigger actions.*

#### 3. Data sources (synthetic/historical data)

##### a. What types and sources of data does the system use?

**Answer:** *The systems primarily use synthetic data, which is artificially generated and contains no personal or sensitive information. This data serves to minimise ethical risks related to privacy protection while enabling controlled coverage of various scenarios. Real-world data is also used to a limited extent, but exclusively under strict conditions: access is restricted to authorised personnel, and all processing is conducted in such a way that it is impossible to identify specific individuals or derive personal information from the data. When sharing or further transferring this data, the principles of data minimisation, anonymisation, and privacy protection are adhered to in accordance with relevant legislation and ethical standards.*

#### 4. Deployment context

##### a. What is the current development stage?

**Answer:** Currently, real data are being captured, tools for generating synthetic data are under development, and synthetic datasets for specific use cases are being created. Selected predictive models (e.g., for road object detection) are trained and evaluated on real and synthetic data.

##### b. What is the system's deployment context?

**Answer:** The systems are developed as standalone predictive modelling components (AI models) and are not deployed as complete operational systems. They do not perform decision-making, provide recommendations, or implement automated interventions. Any integration into downstream applications is outside the scope of the system definition.

##### c. What level of human oversight exists (human-in-the-loop, monitoring), and who is accountable for outcomes?

**Answer:** A human-in-the-loop approach is applied throughout development and evaluation. Human experts are responsible for:

- dataset selection and curation,
- supervision of model training,
- validation and performance assessment, and
- interpretation of model outputs where applicable.

The system produces only predictive outputs, therefore, accountability for any downstream use, interpretation, or decision-making based on these outputs remains with human operators or external systems.

## WP5 - STEP 2: Preliminary identification and mapping of ethical risks of the WP 5 Decision-support system for activities in smart buildings and traffic safety

Based on the system description, the following ethical risks are considered relevant for the development and deployment of the decision-support systems in smart buildings and traffic safety.

### 1. Privacy, data protection, data quality, provenance, and synthetic data risks

The system primarily uses image data which may contain personal data, including identifiable persons, vehicle registration details, and location information. In smart building contexts, image-based monitoring may also capture sensitive information about occupants, visitors, routines, and movement patterns. This can create risks related to data processing, excessive data collection, lack of purpose limitation, and insufficient data minimization. Where synthetic data are derived from real data, there is also a risk that re-identification could remain possible if the generation process is not properly controlled. Real captured data and synthetic data may introduce risks related to provenance tracking, dataset integrity, and representativeness. If the origin, annotation quality, and generation process are not well documented, it may be difficult to assess reliability or reproduce results. Clear traceability between real and synthetic data is therefore needed.

#### Questions:

**A: Did you consider the privacy and data protection implications of data collected, generated or processed over the course of the AI system's life cycle?**

**Answer:** *Yes. Privacy and data protection have been considered across the AI system lifecycle. The system uses primarily synthetic data to minimize reliance on personal data. Where real data is used, it is anonymized or pseudonymized and handled under strict safeguards.*

*All processing complies with GDPR, with clearly defined controller/processor roles, a valid legal basis, and application of data minimization and purpose limitation principles. Dataset creation and use are monitored to mitigate re-identification risks and ensure ongoing compliance.*

**B: Is there any Data Management process implemented on the collection, generation, protection, anonymisation/pseudonimisation, storage, destruction of data?**

**Answer:** *Yes, as described in Data Management Plan (D1.3\_DataManagementPlan\_consolidated version\_final\_v1.0.docx). A data management process focuses primarily on:*

- *controlled handling of real datasets (where used), and*
- *management and documentation of synthetic data generation pipelines.*

*Given the limited use of personal data, the requirements for anonymisation are minimal.*

**C: Did you implement relevant standards or principles (like FAIR Data Principles) for data lifecycle management?**

**Answer:** *The FAIR principles are the key part of the Data Management Plan, which can be found here: D1.3\_DataManagementPlan\_consolidated version\_final\_v1.0.docx.*

**D: Did you put in place any of the following measures, some of which are mandatory under the GDPR?**

**Answer:**

- *Data Protection Impact Assessment (DPIA);*
- *Designate a Data Protection Officer (DPO) early in development;*
- *Oversight mechanisms for data processing (including limiting access to qualified personnel, mechanisms for logging data access and making modifications);*
- *Measures to achieve privacy-by-design and default (e.g. encryption, pseudonymization, aggregation, anonymisation);*

*Partially, where applicable:*

- *Privacy-by-design principles are considered, primarily through minimisation of real personal data usage and preference for synthetic data.*
- *Access to any real datasets is restricted to authorised personnel.*

*Formal GDPR measures (e.g., DPIA) are considered only if real personal data processing exceeds minimal scope.*

**E: Did you put in place measures to ensure the traceability of the data? To what extent can you trace back which data was used by the AI system to make a certain decision/recommendation?**

**Answer:** *Yes. Data provenance is documented at the dataset level, including:*

- *origin of real data (if used),*
- *synthetic data generation processes, and*
- *dataset composition used for model training and evaluation.*

## **2. Technical robustness and safety risks**

Because the system is intended for decision support in safety-relevant contexts, errors in detection or recognition may have operational consequences. Risks include false positives, false negatives, unstable performance under changing environmental conditions, and limited robustness. Synthetic data generation may improve coverage, but it may also introduce

unrealistic artefacts or distribution shifts that reduce performance in real-world settings. If the system is used to support human decisions in operational environments, inaccurate outputs may lead to inappropriate interventions or missed safety incidents.

**Questions:**

**F: Did you put in place a process to measure and assess risks?**

*Answer: Yes. Model evaluation includes performance metrics, validation across datasets, and analysis of prediction errors.*

**G: Did you put measures in place to ensure the integrity, robustness and overall security of the AI system against potential attacks over its lifecycle?**

*Answer: Yes. Standard practices are applied, including validation on diverse datasets, controlled training procedures, and basic security measures for data and model handling.*

**H: Could the AI system have adversarial, critical or damaging effects (e.g. to human or societal safety) in case of risks or threats such as design or technical faults, defects, outages, attacks, misuse, inappropriate or malicious use?**

*Answer: Not yet applicable – WP08.*

**I: Did you identify the possible threats to the AI system (design faults, technical faults, environmental threats) and the possible consequences?**

*Answer: Yes. Identified risks relate to reduced prediction accuracy due to data limitations, domain shifts, or model limitations.*

**J: Did you put in place a series of steps to monitor, and document the AI system's accuracy?**

*Answer: Yes. Model performance can be systematically evaluated and documented using standard metrics.*

**K: Could a low level of accuracy of the AI system result in critical, adversarial or damaging consequences?**

*Answer: Not yet applicable – WP08*

**L: Did you put processes in place to ensure that the level of accuracy of the AI system to be expected by end-users and/or subjects is properly communicated?**

*Answer: Not yet applicable – WP08*

### 3. Human oversight risks

Although the system is not autonomous, there is still a risk that human operators may rely too heavily on automated recommendations or alerts. In decision-support settings, especially where outputs are frequent or seemingly reliable, users may defer to the system even when contextual judgment is required. This is particularly important in traffic safety and smart building monitoring, where false confidence in the system could affect safety-critical decisions. Human-in-the-loop oversight therefore needs to remain substantive.

#### Questions:

**M: Could the AI system affect human autonomy by generating over-reliance by end-users? If yes, did you put in place procedures to avoid that end-users over-rely on the AI system?**

*Answer: Not yet applicable – WP08*

**N: Could the AI system affect human autonomy by interfering with the end-user's decision-making process in any other unintended and undesirable way?**

*Answer: Not yet applicable – WP08*

*Given that the system is strictly limited to predictive outputs without recommendations or decisions, human oversight risks are not directly applicable at the system level.*

*Any potential over-reliance or decision-making impact arises only in external systems or applications, which are not the scope of this work.*

### 4. Misuse and secondary use risks

Systems capable of object detection and human activity recognition can potentially be repurposed for other uses beyond the original research scope. In smart building environments, this may raise concerns about excessive monitoring of occupants. In traffic safety contexts, the same technology could be used for enforcement or profiling purposes beyond the intended decision-support function. This creates a need for strict use limitation, access control, and governance over downstream deployment.

#### Questions:

**O: Have you considered purpose limitations measures?**

*Answer: Yes. The system is explicitly limited to predictive modelling. The operator will not be able to change the model's intended use.*

**Q: Did you assess the risk of possible malicious use, misuse or inappropriate use of the AI system?**

*Answer: Not yet applicable – WP08*

**R: Did you define safety criticality levels (e.g. related to human integrity) of the possible consequences of faults or misuse of the AI system?**

**Answer:** *Not yet applicable – WP08*

### **WP5 - STEP 3: Ethics review outcomes and recommendations for the WP5 Decision-support system for activities in smart buildings and traffic safety**

In Phase 3, the preliminary findings from the curated ALTAI list are formalised into an ethics response for the WP5 decision-support system. This stage records the current status within the ethics framework and a set of recommendations.

For privacy and data governance, the project follows a privacy-by-design approach. The use of synthetic data is prioritised wherever possible to reduce exposure to personal data from the outset. Where real data are required, access is restricted to authorised personnel, and data handling is supported by documented provenance, controlled storage, and lifecycle management. This ensures that the origin, use, and transformation of datasets can be traced reliably, and is supporting compliance with data minimisation, purpose limitation, and security requirements.

For technical robustness and safety, the system is developed and tested through systematic validation across diverse datasets and operating conditions. Performance is monitored using standard metrics, with particular attention to accuracy, robustness, generalisation, and error patterns. The use of synthetic data is controlled so that it improves coverage of edge cases without introducing unrealistic artefacts or harmful distribution shifts.

For misuse and secondary use, the system is bound by strict purpose limitation. Its intended function is restricted to predictive modelling, and governance controls are intended to prevent repurposing outside the approved research scope. Any change in use case, deployment context, or access structure should trigger a new review before further use is done.

For accountability, responsibilities are assigned across the project to ensure that safeguards are maintained over time. Dataset approval, model validation, access control, incident response, and reassessment should be linked to named roles, with documented audit trails and review checkpoints. This strengthens traceability and supports internal oversight.

### **Recommendations**

It is recommended that the project maintain continuous engagement with the ethics team and the external ethics advisors throughout the full lifecycle of the WP5 decision-support system. Ethical review should not be treated as a one-time exercise, but as an ongoing process that accompanies the development, testing, validation, and any future deployment or revision of the system. Regular exchanges should be done so that emerging issues, technical updates, dataset

changes, or shifts in the research direction can be discussed early and addressed before they become problematic.

This iterative approach is especially important because the system is still expected to evolve. As the work progresses, the architecture, data sources, model behaviour, operational context, and intended use may all change. Any such changes should be communicated to the ethics experts and advisors so that the relevant ethical considerations can be reassessed and, where necessary, the mitigation measures can be delivered. The same applies if new functionality is added, if the scope of the research is expanded, if the deployment context becomes more operational, or if the system begins to rely on different categories of data.

The project team should therefore keep ethics involvement active across the whole process. Such approach will help ensure that ethical safeguards are in place.

## Annex 3 – WP6

### WP6 - STEP 1: System Characterisation and Contextual Understanding

#### 1. Capabilities

- a. What type of system will you develop? (e.g., decision-support tool, automation system, predictive model, data analytics platform, generative AI system, etc.)

**Answer:** We are developing an AI-augmented decision-support tool for automating and enhancing the software development lifecycle. The system utilizes open-weight LLMs (Mistral family) fine-tuned via Unsloth. It acts as an AI assistant to support software modeling, the transformation of analysis artifacts into design assets, and code maintenance. It evaluates complex development contexts and generates proposals for software engineers.

- b. In which areas does the system make decisions entirely autonomously?

**Answer:** None. The system is strictly designed as a decision-support and automation assistant for software engineers. It does not execute autonomous changes to production code or architecture without explicit human validation.

#### 2. Intended use

- a. In what scenarios will the system be used?

**Answer:** It will be used within secure software development environments to assist IT professionals. Scenarios include analyzing software requirements, generating architecture design proposals, monitoring systems, and researching secure approaches to handle source code with AI tools.

#### 3. Data sources (synthetic/historical data)

- a. What types and sources of data does the system use?

**Answer:** During R&D and Fine-Tuning: The models are trained on synthetic datasets and anonymized software development artifacts to ensure the security of sensitive source code. During Production/Deployment: The system will process software analysis artifacts, architectural documents, and relevant source code repositories within a strictly controlled environment.

#### 4. Deployment context

- a. What is the current development stage?

**Answer:** We are currently in the R&D and fine-tuning stage (evaluating models against the HELM framework using 7 core metrics: accuracy, calibration, robustness, fairness, bias, toxicity, and efficiency).

### i. What is the system's deployment context?

**Answer:** *The system will be deployed in a secure Cloud environment on virtual machines. The environment will strictly adhere to EU security requirements, including data residency within the EU. All access and communication with the LLM models will be via encrypted network connections, and no data will be routed to third-party APIs.*

### b. What level of human oversight exists (human-in-the-loop, monitoring), and who is accountable for outcomes?

**Answer:** *Human-in-the-Loop: Strict human oversight is architecturally enforced. The system generates proposals, but the implementation requires manual approval from the software engineer.*

*Monitoring: A secondary, superior "Judging LLM model" will be deployed to continuously monitor the daily inputs and outputs of the production model.*

*Accountability: The human software engineer and the deploying organization remain fully accountable for all final code and architectural decisions.*

## WP6 -STEP 2: Preliminary identification and mapping of ethical risks

Based on the system description, the relevant ALTAI dimensions for this AI-augmented software development assistant are human agency and oversight, technical robustness and safety, privacy and data governance, transparency, and accountability. The system is not intended to make decisions about individuals, allocate rights, or perform autonomous actions in production, and is therefore not associated with high-risk fundamental rights impacts.

In addition, the system is evaluated under the HELM framework using the metrics accuracy, calibration, robustness, fairness, bias, toxicity, and efficiency, which further reduces potential risks and systematically assess model behaviour across performance and safety dimensions.

### 1. Human agency and oversight

The system is designed as a support tool for software engineers and cannot execute autonomous changes. Final decisions remain with the human engineer, and this preserves meaningful human control over software design and maintenance. The main residual risk can be over-reliance on generated proposals, but the system itself does not undermine human agency by design.

### 2. Technical robustness and safety

The relevant risk is limited to the quality and reliability of the generated suggestions, including possible errors, hallucinations, or insecure recommendations. These are technical quality risks, not direct autonomous safety risks. Among these there is also a risk that generated proposals or refactoring suggestions may weaken security controls, introduce vulnerabilities, or normalize

unsafe development patterns. As a mitigation measure, the secure cloud architecture and prohibition on third-party APIs will reduce exposure.

### 3. Privacy and data governance

The system is not designed to process personal or sensitive personal data. In the R&D phase, it relies on synthetic datasets and anonymised software development artefacts, and in deployment it operates in a controlled environment with EU data residency and no third-party APIs.

### 4. Transparency

The system outputs are reviewable proposals rather than binding decisions. Nevertheless, the system should still remain transparent about which inputs were used, which artefacts were analysed, and how a given suggestion was produced. So, the system does not pose a transparency risk, provided that its outputs remain traceable and clearly labelled as machine-generated support.

### 5. Accountability

The system does not create an accountability gap, because responsibility for final code and architecture decisions remains with the human software engineer and the deploying organisation. The AI assistant may propose, summarise, or monitor, but it does not approve or deploy changes on its own. This means that accountability is structurally preserved, and the system does not shift responsibility away from the human decision-maker.

### 6. Intellectual property and licensing risks

Software repositories and architectural documents may be governed by internal policies or external licenses. If the system is trained or used on code whose licensing terms are restrictive, there may be legal and operational risks related to derivative outputs, redistribution, or code reuse. Generated code or proposals may also unintentionally resemble protected or licensed material too closely, which could create compliance concerns. This is particularly relevant when using LLMs for code generation or transformation, because output provenance and originality are not always easy to establish.

### Recommendations:

The system should remain bound to advisory use, with mandatory human validation before any code or architectural change is applied. Attention should be given to intellectual property and licensing compliance: the team should verify the provenance and licence status of any source code, architectural artefacts, training data, and generated outputs, and should avoid training or deployment on material whose reuse, redistribution, or derivative transformation may be restricted. Any substantial changes to the system, the data pipeline, or the deployment setup should be communicated to the Ethics Team, together with regular updates on the progress of HELM implementation and evaluation results, so that the ethics evaluation remains updated and aligned with the evolving system.

## Annex 4 – WP7

### WP7 - STEP 1: System Characterisation and Contextual Understanding

#### 1. Capabilities

- a. What type of system will you develop? (e.g., decision-support tool, automation system, predictive model, data analytics platform, generative AI system, etc.)

**Answer:** *We are developing an AI-powered decision-support assistant designed for the technical and administrative pre-processing of insurance claim documentation. The system prepares structured, evidence-based summaries to support human claims handlers in their assessment.*

*The system supports:*

- *Intake and processing of claim inputs.*
- *Automated preparation and structuring of case information.*
- *Evidence-based assessment support (extraction of facts from documents).*
- *Rule-based validation and consistency checks.*
- *Preparation of a non-binding draft assessments and supporting rationale for subsequent human review.*
- *Full explainability and traceability for review and audit purposes.*

*In practice, the system automates the operational workload across the full case flow, but the final legal and business decision remains exclusively with an authorized human reviewer (who can review, modify or disregard system output).*

- b. In which areas does the system make decisions entirely autonomously?

**Answer:** *The system does not make autonomous decisions with legal effects. It performs exclusively automated procedural and technical tasks within the operational workflow:*

- **Technical triage:** *Quality and completeness checks of input data (identifying missing documents).*
- **Case routing:** *Proposes assignment of the case to the correct next state (for example: continue processing, request additional information, escalate to human review, technical failure handling).*

- **Preparation of structured intermediate outputs:** Generating case-level summaries and non-binding recommendations to serve as a baseline for the human expert.

*Human-only area:*

- *Final legal/business approval of the claim outcome (granting or denying the claim).*

*All outputs are subject to human review, and qualified personnel retain full authority and responsibility for the final decision.*

## 2. Intended use

### a. In what scenarios will the system be used?

**Answer:** *The system is intended for internal use by insurance claim teams.*

*Main scenarios:*

- *Processing incoming claim materials and case records.*
- *Building a structured and comprehensive review package for each case.*
- *Detecting missing evidence, data inconsistencies, or extraction uncertainties.*
- *Producing structured outputs and non-binding recommendations to support human review and decision-making*

*Flagging and escalating potentially high-risk or ambiguous cases for enhanced human assessment. The intended operating model is built on high degree of workflow automation, but all outputs are subject to human review and final decisions are made exclusively by qualified human reviewers, who are required to assess and validate system outputs before any outcome is confirmed.*

## 3. Data sources (synthetic/historical data)

### a. What types and sources of data does the system use?

**Answer:** *Primary external/business data sources (used for real case processing):*

- *Claim documents submitted in the claims process (for example: claim forms, medical reports, supporting incident documentation).*
- *Structured claim/case records provided by insurer core systems.*

- *Policy and product documentation relevant to entitlement assessment (terms, conditions, limits, exclusions, and applicable process rules).*
- *Historical claim records made available by the insurer for contextual support.*

*Supplementary internal project datasets (non-primary for live decisions):*

- *Synthetic datasets created by the project team for development, testing, and controlled evaluation.*

*Data handling principles:*

*The system is designed in accordance with the principles of the GDPR, including data minimization, purpose limitation, and appropriate technical and organizational safeguards.*

- **Data minimization and preprocessing:** *Personal data is limited to what is necessary for the specific processing purpose. As part of preprocessing, automated modules apply anonymization/pseudonymization to reduce direct identifiability before data is used by AI components.*
- **Sensitive Data Handling:** *Medical data (diagnoses, ICD codes) are processed in a secure environment, with models receiving only the necessary context decoupled from the individual's identity.*
- **Use of Synthetic Data:** *For development and testing, synthetic datasets are prioritized to eliminate risks associated with real personal data.*
- **Traceability and Logging:** *All processing steps, including extraction confidence scores, are comprehensively logged, ensuring full accountability and a technical audit trail.*
- **Access control and security:** *Access to data and system outputs is restricted to authorized personnel based on role and necessity, supported by appropriate technical and organizational measures.*

## 4. Deployment context

### a. What is the current development stage?

**Answer:** *The system is in PoC development stage.*

*Current maturity (realistic status):*

- *Core document and case-processing foundations are implemented.*

- *Some recommendation-flow elements are running, while several planned workflow components are still in progress or not yet implemented.*
- *A substantial part of integration and monitoring scope remains open in the current plan.*

**b. What is the system's deployment context?**

**Answer:** *PoC deployment context:*

- *Controlled development and integration environments.*
- *Internal testing and validation with project and business stakeholders.*
- *Limited-scope PoC operation for selected scenarios only*

**c. What level of human oversight exists (human-in-the-loop, monitoring), and who is accountable for outcomes?**

**Answer:** *Human oversight:*

- *Human-in-the-loop is mandatory for final case closure. Human review is a substantive and integral part of the workflow, supported by system-generated citations and traceable links to source documents to enable informed verification.*
- *The system identifies and flags cases that may require additional review based on detected indicators such as uncertainty, data gaps, evidence inconsistencies, or potential risk conditions, thereby supporting appropriate human intervention.*
- *Human reviewers are required to assess system outputs and retain full authority to confirm, modify, or reject any recommendations.*  
*Accountability:*
- *Final accountability remains with authorized insurer decision-makers.*
- *The AI system provides analytical support, structured outputs, and traceability mechanisms, while legal and business responsibility stays with human authorities.*

## WP7 -STEP 2: Preliminary identification and mapping of ethical risks

The system is an AI decision-support and workflow-assistance tool for insurance claims. It automates technical preprocessing tasks such as document classification, extraction, validation, and case structuring, and may propose routing or recommendations, but it does not make final legal or business decisions. Final claim approval or denial always remains with a human reviewer.

### 1. Privacy and data protection risks

During Production/Deployment, the system will process sensitive data and personal data contained in claim files, client communications, accident documentation, and historical claim records. In some cases, this may include special category data such as health information, as well as other sensitive contextual data relevant to liability or entitlement assessment. Even though the system is designed to use pseudonymised or anonymised data for training and testing, risks remain if data minimisation is incomplete, if re-identification becomes possible through combined context, or if operational data are exposed through logs, summaries, or model outputs.

The use of LLM-based processing introduces privacy risks because models may retain, infer, or reproduce sensitive content from inputs if safeguards are insufficient. The planned deployment in a secure cloud environment within the EU is a positive mitigation factor, but privacy risk remains relevant throughout the lifecycle, especially where claims documents may contain extensive personal narratives, third-party references, or medical details unrelated to the minimal processing purpose.

#### Questions:

**A:** Is there any Data Management process implemented on the protection, anonymisation / pseudonimisation, storage, destruction of data?

**ACE:** YES. A dedicated Data Management Plan (DMP) has been established. At the outset of the process, the categories of data to be used were clearly identified. As the potential involvement of personal data was recognized, appropriate mitigation strategies were adopted. Initially, synthetic data were used for training purposes. From a technical and accuracy perspective, it was determined that further improvements to the model's performance would require a GDPR-compliant approach. Various use cases were therefore tested to assess the legal basis for such processing. The outcome of this assessment indicates that training can be effectively conducted using anonymized datasets. For the current use cases, anonymized datasets will be provided by the data handler, eliminating the need to process personal data directly. Ensuring compliance remains a high priority, which is why the system has been designed accordingly. It includes a specialized anonymization module developed within WP07 that removes personal identifiers (PII) before any data is processed by the LLM core.

**GTI:** YES (TRL4-TRL6). GTI enforces a strict external preprocessing boundary: depersonalisation, anonymisation, and encryption are executed outside the InnovAite/GNOTHEIA ecosystem before any intake. No real-world personal data is used in LLM inference. In line with the data provisioning

*protocol declarations (PIC protocol wording under the cooperation framework), GTI applies a very reduced authorised-person list for this data regime, and those authorised persons can produce only deterministic aggregated statistical outputs and GDPR-compliant statistical reports for project KPI evaluation and SYDAFA hinting/calibration. Raw or non-aggregated records are excluded from AI/ML/LLM training, validation, benchmarking, and inference use. Any non-compliant data is subject to immediate deletion from any data sources and research ecosystem.*

**B: Did you implement relevant standards or principles (like FAIR Data Principles, data minimization, etc.) for data lifecycle management?**

**ACE:** YES. We strictly follow the data minimisation principle. Only data essential for claim assessment (e.g., ICD codes, treatment duration) is extracted, while unrelated personal narratives are filtered out.

**GTI:** YES (TRL4-TRL6). We apply data minimisation by design, purpose limitation, and FAIR-aligned metadata. Inside InnovAlte/GNOTHEIA, processing is limited to GDPR-ethically preprocessed inputs and synthetic artefacts; no direct handling of raw personal data is allowed in LLM inference paths.

**C: Did you put in place any of the following measures, some of which are mandatory under the GDPR?**

○ **Data Protection Impact Assessment (DPIA);**

**ACE:** YES. In progress. A DPIA template and processing-impact documentation are prepared for the PoC, but the formal review and sign-off have not yet been finalized, given the current stage of development described above.

**GTI:** YES (TRL4-TRL5 now, TRL6 sign-off target). DPIA preparation is in progress for the WP7 PoC and planned production path. Risk scenarios, processing records, and mitigation controls are documented; formal sign-off is required before production-grade rollout.

○ **Designate a Data Protection Officer (DPO) early in development;**

*Partially: The project does not yet have a formally designated DPO at project level however a DPO exists at the organisational level (Asseco/GTI), and data protection topics are actively being consulted through governance structure.*

○ **Oversight mechanisms for data processing (including limiting access to qualified personnel, mechanisms for logging data access and making modifications);**

**ACE:** YES. The system implements role-based access control (RBAC) and comprehensive logging of all data access and modifications.

**GTI:** YES (TRL4-TRL6). Oversight includes RBAC, least-privilege access, environment segregation, immutable audit logging, and change traceability for data transformations and model/prompt artifacts.

○ **Measures to achieve privacy-by-design and default (e.g. encryption, pseudonymization, aggregation, anonymisation);**

**ACE:** YES. Measures include end-to-end encryption, irreversible tokenization of identifiers, and local-first processing before any cloud interaction.

**GTI:** YES (TRL4-TRL6). Privacy-by-design controls are applied with a hard boundary: anonymisation and encryption are performed outside the InnovAite ecosystem, and only compliant preprocessed inputs enter system components. No real-world personal data are processed in LLM inference paths in the WP7 research setup.

**D: Did you put in place measures to ensure the traceability of the data? To what extent can you trace back which data was used by the AI system to make a certain decision/recommendation?**

**ACE:** YES, partially. The current PoC already carries trace IDs, routing telemetry, artifact lineage, and versioned processing metadata. Full end-to-end recommendation traceability, including final decision reasoning, prompt/version provenance, and cited evidence, is the target design and is only partially implemented at the current stage.

**GTI:** YES (TRL4-TRL6). Data traceability is implemented according to GNOTHEIA module allocation and strict boundary rules. External provider-side preprocessing (depersonalisation, anonymisation, and encryption) is outside InnovAite and enters through SERVICE INTEGRATION only as GDPR-compliant preprocessed intake. Within GNOTHEIA KNOWLEDGE SPACE, POLYCONTEXT PROCESSOR maintains lineage for preprocessed/synthetic artefacts: evidence pointer IDs, extraction lineage from KNOWLEDGE EXTRACTOR, rule-evaluation trace from RULES/ONTOLOGIES, model/prompt version provenance, DETERMINISTIC GUARDRAILS validation status, confidence/escalation flags, and human override/approval logs in POLYCONTEXT JOURNAL. No trace chain includes raw real-world personal data in LLM inference paths. Full end-to-end evidencing from provider protocol declaration to human-supported output is the TRL5-TRL6 hardening target.

## 2. Fairness, bias, and non-discrimination risks

Insurance claim assessment is a domain in which biased outputs can have direct material consequences. The system may inherit historical patterns from past claims data, policy wording, or prior decision behaviour that reflect structural bias, inconsistent practices, or unequal treatment of certain categories of claimants. This is especially important where the model is trained or fine-tuned on synthetic knowledge sets derived from historical data that may already encode past bias.

There is a risk that the system may systematically perform differently across claim types, medical conditions, demographic groups, legal categories, or language styles. For example, claimants whose documentation is less standardised, less complete, or written in non-standard languages. Such risks make fairness monitoring and group-wise performance evaluation highly relevant.

### Questions:

**E: Did you establish a strategy or a set of procedures to avoid creating or reinforcing unfair bias in the AI system, both regarding the use of input data as well as for the algorithm design?**

**ACE:** YES. In the current PoC, a structured procedural approach is implemented to mitigate bias

*and ensure fairness. Deterministic rules take precedence over LLM outputs, and unclear or low-confidence cases are escalated to human review (human-in-the-loop).*

*Bias is monitored through comparative analysis across operational segments (e.g., document type, channel, product, routing outcomes), supported by stratified testing (golden set), error analysis based on reason\_codes, and post-audit sampling of decisions.*

*The PoC also defines bias signals (e.g., differences in needs\_hitl rates, OCR/parsing success, or decision distributions across segments), with indicative thresholds (~15–20%) triggering root-cause analysis.*

*Bias mitigation is implemented procedurally rather than via model retraining, including stricter rule-based guardrails, improved input data quality (OCR/parsing), enforced citation requirements, and adaptive routing (increased human oversight in higher-risk segments).*

*These measures ensure auditability, transparency, and controlled deployment in line with PoC constraints.*

**GTI:** YES. (TRL4-TRL7 roadmap). The bias-control strategy is a cross-cutting layer aligned with the GNOTHEIA ecosystem allocation and WP7 BIAS-fight baseline. Input-side controls: external ethical preprocessing boundary + minimisation, and synthetic-data-first evaluation/training datasets via SYDAFA. Algorithm/process controls: LLM is used for extraction/decision support only; no autonomous final claim decision; DETERMINISTIC GUARDRAILS and SBVR/ontology-governed gates enforce policy constraints; low-confidence/high-impact cases are escalated to mandatory human-in-the-loop review. Monitoring controls: GNOTHEIA ETHICAL MONITOR executes HELM-aligned lifecycle fairness evaluation; EXPLAINABILITY MONITOR supports post-hoc analysis; segment-level disparity tracking is used, while Shapley attribution is applied to core claim-management gates and not to red-flag deterministic entry gates.

## **F: Did you consider diversity and representativeness of end-users and/or subjects in the data?**

### **o Did you test for specific target groups or problematic use cases?**

**ACE:** YES. Testing includes diverse types of medical reports from various healthcare providers to ensure linguistic and structural robustness.

**GTI:** YES (TRL4-TRL6). We test specific target groups and problematic use cases using synthetic datasets only. SYDAFA uses hints derived exclusively from aggregated GDPR-compliant statistical benchmark profiles; no real-world record-level data are injected into the SYDAFA generator. No raw or record-level real-world personal data is used in LLM inference paths.

### **o Did you research and use publicly available technical tools, that are state-of-the-art, to improve your understanding of the data, model and performance?**

**ACE:** YES. We utilize fairness-monitoring frameworks and collaborate with DFKI and UNIBA to evaluate model performance across demographic segments.

**GTI:** YES (TRL4-TRL6). State-of-the-art tooling includes HELM-aligned evaluation through GNOTHEIA ETHICAL MONITOR, plus structured quality/fairness harnesses for extraction consistency, error profiling, and segment disparity checks. These are combined with deterministic gate validation and evidence-trace checks.

- **Did you assess and put in place processes to test and monitor for potential biases during the entire lifecycle of the AI system (e.g. biases due to possible limitations stemming from the composition of the used data sets (lack of diversity, non-representativeness))?**

**ACE:** YES. Continuous monitoring is part of the WP07.5 deliverable (Monitoring Protocol)

**GTI:** YES (TRL4 now; TRL5-TRL7 continuous). Bias testing and monitoring are lifecycle controls: periodic HELM-based evaluation runs, segment-level disparity tracking, drift checks, incident-triggered re-evaluation, and corrective-loop governance. Operationally, this is coupled with HITL escalation and deterministic red-flag gate controls.

**G: Did you ensure a mechanism that allows for the flagging of issues related to bias, discrimination or poor performance of the AI system?**

**ACE:** YES. Partially. In the current PoC, mechanisms are in place to flag issues related to bias, discrimination, and system performance through integrated workflow controls.

Issues are identified via rule-based checks, inconsistency detection, and confidence indicators (e.g., LOW\_CONFIDENCE, RAG\_CITATION\_MISSING), and are flagged and routed using predefined states (e.g., needs\_hitl, needs\_info), ensuring escalation to human experts.

Additional monitoring is performed through comparative analysis across segments (e.g., product, channel, document type), supported by error analysis (reason\_codes) and post-audit sampling. Significant deviations (~15–20%) trigger further investigation.

All issues are traceable via audit logs, ensuring transparency and auditability. Formal governance processes are planned for later phases (WP08).

**GTI:** YES (TRL5-TRL7 target). A structured flagging mechanism is defined: users/reviewers can mark bias/discrimination/poor-performance cases; each case is triaged, logged, and linked to corrective actions with ownership and SLA.

**H: Did you establish clear steps and ways of communicating on how and to whom such issues can be raised?**

**ACE:** YES. Partially. In the current PoC, clear procedural steps for raising and handling issues are implemented through system workflows rather than formal governance structures.

Issues are primarily raised via automated routing mechanisms (e.g., needs\_hitl), which escalate unclear or low-confidence cases to human experts. Additional issues are identified through audit

*logs, reason\_codes analysis, bias monitoring (including threshold-based triggers), and post-audit sampling.*

*These mechanisms define how issues are detected and escalated within the system. However, formal communication procedures (e.g., defined roles, responsibilities, and escalation channels) are not yet fully specified and are planned to be established in later stages of the project (WP08).*

**GTI:** YES (TRL5-TRL7 target). Escalation and communication paths are defined by role (operator -> supervisor -> compliance/privacy -> technical owner). Communication templates specify what to report, required evidence, severity classification, and expected response timelines/SLA for bias, discrimination, or poor-performance incidents.

### **I: Did you establish mechanisms to ensure fairness in your AI system?**

**ACE:** YES. We implement rule-based guardrails that ensure identical medical conditions result in consistent recommendations regardless of the claimant's background.

**GTI:** YES (TRL4-TRL7). Fairness is implemented as a multi-component control chain across the GNOTHEIA ecosystem. Core controls include SBVR/ontology-governed deterministic gates (GNOTHEIA RULES EXTRACTOR + RULES REPOSITORY + ONTOLOGIES REPOSITORY), runtime policy enforcement via DETERMINISTIC GUARDRAILS in the POLYCONTEXT PROCESSOR flow, and mandatory HITL review in GNOTHEIA AGENT for high-impact/low-confidence cases. Fairness evaluation is monitored through GNOTHEIA ETHICAL MONITOR using HELM-aligned checks (including segment-level disparity and consistency checks), with explainability support from GNOTHEIA EXPLAINABILITY MONITOR. Test scenarios are generated through SYDAFA from aggregated GDPR-compliant benchmark hints only; no real-world record-level data are injected into SYDAFA and no raw/record-level real-world personal data are used in LLM inference paths.

### **J: Did you ensure a quantitative analysis or metrics to measure and test the applied definition of fairness?**

**ACE:** NO. for demographic-parity-style metrics at this stage. The current PoC does not rely on demographic fairness metrics such as Demographic Parity or Equalized Odds; instead, it uses segment-based bias signals and post-audit analysis.

**GTI:** YES, (TRL4-TRL6). We apply quantitative fairness testing through a HELM-aligned evaluation protocol in GNOTHEIA ETHICAL MONITOR, combined with deterministic gate analytics. Current fairness checks are operational-segment based (not demographic-parity-only): (1) decision/recommendation change rate across controlled counterfactual pairs, (2) escalation-rate disparity, (3) extraction error disparity, (4) evidence-citation consistency disparity, and (5) confidence-gap disparity by segment. Execution chain: SYDAFA generates synthetic counterfactual test sets from aggregated GDPR-compliant benchmark hints only; POLYCONTEXT PROCESSOR + LLM MODEL produce outputs; DETERMINISTIC GUARDRAILS enforce rule constraints; ETHICAL MONITOR computes HELM-aligned fairness/robustness scores; EXPLAINABILITY MONITOR supports post-hoc analysis for anomalous disparities; POLYCONTEXT

*JOURNAL + PANOPTUS store auditable metric history and incidents. No real-world record-level data are injected into SYDAFA, and no raw/record-level personal data are used in LLM inference paths.*

**K: Did you consider a mechanism to include the participation of the widest range of possible stakeholders in the AI system’s design and development?**

**ACE:** YES. The consortium includes end-users as a Regional Government (TSK) and Academic researchers (TNUAD, UNIBA).

**GTI:** YES. (TRL4-TRL7). Within the InnovAite WP7 consortium, stakeholder participation for AI design/development is organised around technical and governance roles: AI/LLM engineering, data engineering, ontology/SBVR rule engineering, synthetic-data design (SYDAFA), evaluation/fairness monitoring (HELM-aligned Ethical Monitor), explainability analysis, QA, and compliance/privacy review. We do not claim operational participation by the claim handler/adjuster as a standing consortium role in this process. A targeted business-review commitment is in place: the data provider's Claim Department Head reviews selected design decisions and the results of polycontext evaluations performed on synthetic test cases before major iteration checkpoints.

### 3. Technical robustness, accuracy, and safety risks

Inaccuracies in the system may cause delays, incorrect recommendations, unnecessary escalations, or improper reliance on flawed intermediate outputs. LLM systems in particular may generate plausible but incorrect summaries, misread document content, or infer unsupported conclusions. This creates a risk of hallucination, omission, or distortion of evidence. The combination of unstructured document processing, extraction, rule evaluation, and summarisation also creates a multi-stage error propagation problem. A mistake in early document classification or extraction may influence the polycontext and downstream recommendation. Even if no final automated decision is made, technical errors can still shape the human review process in practice. The planned use of a secondary “judging” model explainability is useful, but it does not remove the need for continuous monitoring, and fail-safe handling of low-confidence outputs.

**Questions:**

**L: Did you consider different types of vulnerabilities and potential entry points for attacks such as: Data poisoning (i.e. manipulation of training data); Model inversion (i.e. infer the model parameters); etc.?**

**ACE:** YES. Partially. The security model is based on a hybrid architecture: pre-anonymisation processing remains on-premises, while only anonymised data may be sent to EU-based cloud LLM services. Token-based and optional mTLS controls are implemented for storage services, while end-to-end application-layer access control is still being completed as part of the PoC hardening.

**GTI:** YES. (TRL4-TRL6). We maintain an explicit threat model covering data poisoning, prompt injection, model inversion/extraction attempts, sensitive-data exfiltration, privilege escalation, and insecure integration endpoints. Security architecture is component-mapped: SERVICE INTEGRATION enforces controlled intake and interface hardening; external ethical preprocessing boundary prevents raw real-world personal data from entering LLM paths; POLYCONTEXT PROCESSOR runs with deterministic gate checks (DETERMINISTIC GUARDRAILS) and confidence-based escalation; RULES/ONTOLOGIES constraints reduce unsafe recommendation drift. Monitoring/observation is performed through PANOPTUS + POLYCONTEXT JOURNAL (audit trail, anomaly/event tracking, incident evidence), while GNOTHEIA ETHICAL MONITOR performs HELM-aligned robustness/fairness checks over synthetic/counterfactual evaluation sets. Glass-box explanatory support is provided via EXPLAINABILITY MONITOR plus evidence-pointer/citation tracing, enabling post-hoc forensic review of model-supported outputs. Controls include RBAC/least-privilege, environment segregation, encryption in transit/at rest, immutable logging, and incident-triggered re-evaluation.

**M: Did you inform end-users and subjects of existing or potential risks?**

**ACE:** N/A – WP08

**GTI:** YES, PARTIALLY (WP7 TRL4-TRL6 scope; WP8+/production follow-up). In the WP7 research phase we do not operate with production end-user roles, so formal end-user risk communication is not fully active yet. However, the GNOTHEIA evidence pack and planned operational governance explicitly prepare this for the following phases: the planned Protocol set (Service Entry, Solution Delivery, Explain Request, Operations Monitoring, Ethical Monitoring, Support Request, Data Access Request, plus Tuning and QA) defines how risk/limitation communication, escalation, and auditability will be executed in WP8 and potential production deployment. For WP7, communication is limited to consortium technical/governance stakeholders and to a targeted review commitment by the data provider, the Claim Department Head, of synthetic polycontext evaluation results.

**N: Did you assess the risk of possible malicious use, misuse or inappropriate use of the AI system?**

**ACE:** N/A – WP08

**GTI:** YES (TRL4-TRL6). We assess malicious use, misuse, and inappropriate-use risks through a structured threat and misuse model aligned to GNOTHEIA deployment strategy. Covered scenarios include prompt injection, adversarial/poisoned inputs, exfiltration attempts, over-reliance on AI suggestions, bypass of deterministic gates, and unauthorised data-access paths. Mitigations are mapped to components and planned protocols: DETERMINISTIC GUARDRAILS + SBVR/ontology constraints, mandatory HITL escalation, PANOPTUS/POLYCONTEXT JOURNAL observability, and incident handling via the planned Support + Operations Monitoring + Data Access protocol chain for later phases.

**O: Did you put in place a series of steps to monitor and document the AI system's accuracy?**

**ACE:** YES. We will measure the "Alignment Rate" between AI recommendations and final human decisions.

**GTI:** YES (TRL4-TRL7, scope-correct for decision support). Accuracy monitoring is implemented as a multi-layer quality process, not as AI vs human decision replacement scoring. Since GNOTHEIA does not issue final claim decisions, we do not treat alignment between recommendations and final decisions as the primary success metric. Primary monitored metrics are: extraction precision/recall and completeness; evidence grounding/citation correctness; rule-constraint consistency at deterministic gates; confidence calibration bands; escalation/override rates; and error taxonomy trends over retest cycles. Monitoring evidence is captured in POLYCONTEXT JOURNAL/PANOPTUS and evaluated through QA + Ethical Monitoring controls (HELM-aligned checks where applicable).

**P: Did you put processes in place to ensure that the level of accuracy of the AI system to be expected by end-users and/or subjects is properly communicated?**

**ACE:** N/A – WP08

**GTI:** PARTIALLY (WP7 TRL4-TRL6; WP8+/production follow-up). In WP7, communication of expected accuracy is targeted to consortium technical/governance roles and documented in evaluation/report artefacts, including limitations, confidence cues, and mandatory escalation conditions. For later phases (WP8 and potential production), formal communication to operational users/subjects is planned through the GNOTHEIA protocol evidence pack (Service Entry, Explain Request, Operations Monitoring, Ethical Monitoring, Support Request, Data Access, Solution Delivery, plus QA/Tuning), with explicit accuracy bands, known limits, and fallback/escalation rules.

#### 4. Transparency and explainability risks

The system aims to provide explainability and auditability, which is important in a claims-processing context. However, explainability and transparency also concerns the limits of the system: users must understand what the system can and cannot infer, where confidence is weak, and when the output should not be relied upon.

##### Questions:

**Q: Did you establish mechanisms to inform users about the purpose, criteria and limitations of the decision(s) generated by the AI system?**

- Did you communicate the benefits of the AI system to users?

**ACE:** N/A – WP08

**GTI:** YES, PARTIALLY (WP7 TRL4-TRL6; WP8+/production follow-up). Q(1) Yes for WP7 internal users: purpose, criteria, and limitations are communicated through evidence-

*pointer traces, SBVR/ontology rule references, confidence cues, and deterministic gate status in evaluation artefacts. Q(2) Benefits are communicated at the technical/governance level (faster evidence synthesis, more consistent rule evaluation, improved auditability), explicitly as decision support only and not as human decision replacement. Formal end-user communication packages are planned for WP8/production via Explain Request, Service Entry, and QA/Solution Delivery protocols.*

- **Did you communicate the technical limitations and potential risks of the AI system to users, such as its level of accuracy and/ or error rates?**

**ACE: N/A – WP08**

**GTI:** YES, PARTIALLY (WP7 TRL4-TRL6; WP8+/production follow-up). Technical limitations and potential risks are documented for consortium use in WP7: error modes, confidence uncertainty, non-deterministic behaviour zones, and mandatory escalation/override conditions. For WP8/production, this is planned to be communicated to operational users through Explain Request, Operations Monitoring, Ethical Monitoring, and Support protocols, including explicit accuracy/error bands and fallback rules.

- **Did you provide appropriate training material and disclaimers to users on how to adequately use the AI system?**

**ACE: N/A – WP08**

**GTI:** YES, PARTIALLY (WP7 TRL4-TRL6; WP8+/production follow-up). In WP7, we provide guidance artefacts and review instructions for consortium experts working with synthetic evaluation cases; these are not yet full production user training packs. Comprehensive training materials, disclaimers, and role-specific SOPs for operational users are planned for WP8/production rollout under Service Entry, Explain Request, and Support protocol governance.

**R: Do you survey the users if they understand the decision(s) of the AI system?**

**ACE: N/A – WP08**

**GTI:** YES, PARTIALLY (WP7 TRL4-TRL6; WP8+/production follow-up). In WP7 we validate interpretability mainly through expert review of synthetic polycontext evaluations (including targeted review by the data provider's Claim Department Head). Formal recurring user-understanding surveys are planned for WP8/production operation under the Explain Request and Operations/Ethical Monitoring protocol workflow.

## 5. Accountability and governance risks

The system is explicitly positioned as decision support, with final accountability remaining with the insurer and its authorised operators. Even so, the layered system architecture creates accountability complexity. If a claim outcome is influenced by a model-generated summary, a

rule-evaluation result, or an extracted document classification, responsibility may become diffused across developers, deployers, and human decision-makers. This creates a governance risk if the division of responsibilities is not formally documented.

To mitigate accountability and governance risks, the insurer could establish a formal governance framework that assigns responsibility for each stage of the operational workflow. This must explicitly inform that all final decisions remain solely with the human decision-maker, and that the AI system serves strictly as a support tool without decision-making authority. Human operators must be formally informed, trained, and periodically reminded of their responsibility for claim outcomes. This includes clear internal policies stating that system-generated summaries, classifications, or recommendations are non-binding and must be assessed before any decision is made. Operators should be required to actively review, validate, and, where necessary, challenge or override system outputs.

**GTI:** *Accountability and governance are treated as explicit architectural controls, not informal assumptions. GNOTHEIA is strictly decision-support: it does not issue final claim decisions, and final responsibility remains with the insurer-side authorised human decision-maker. For WP7 (TRL4-TRL6), governance evidence is maintained through POLYCONTEXT JOURNAL and PANOPTUS audit trails, deterministic SBVR/ontology gate enforcement, documented escalation/override traces, and role-scoped technical/compliance review loops (including targeted review of synthetic polycontext evaluation outputs by the data provider's Claim Department Head). For WP8+/potential production deployment, the accountability model is formalised via the planned protocol evidence pack (Tuning, QA, Solution Delivery, Service Entry, Explain Request, Operations Monitoring, Ethical Monitoring, Support Request, Data Access Request) with clear responsibility boundaries, escalation paths, rollback authority, and auditable decision-support non-binding status. This governance stance remains fully aligned with the research project data boundary: no raw/record-level real-world personal data in LLM inference paths and SYDAFA hinting from aggregated GDPR-compliant statistical benchmarks only.*

## WP7 - STEP 3: Ethics review outcomes and recommendations

In Phase 3, the findings from the curated ALTAI list are formalised into an ethics response for the WP7 AI-powered decision-support assistant. This stage records the status within the ethics framework and a set of recommendations.

For privacy and data governance, the project established a Data Management Plan and applies a privacy-by-design approach. Training and testing are based on synthetic or anonymised datasets wherever possible, and the system incorporates an anonymisation module that removes personal identifiers before information is processed by the LLM. Data traceability is supported through processing metadata, audit trails, and version control. The Data Protection Impact Assessment remains in progress.

Fairness monitoring, bias mitigation, and human oversight are addressed.

For technical robustness and safety, the system is designed as a decision-support tool and does not make autonomous claim decisions. Security safeguards are in place, and system performance is monitored.

Transparency and explainability: the project provides evidence-based outputs. During the current research phase, the explainability mechanisms are intended for consortium partners and technical reviewers. Formal communication of the system's capabilities, limitations, expected accuracy, and appropriate use to operational users, together with user training materials and disclaimers, are planned for further project phases.

Accountability: the system is designed as a decision-support tool, with final responsibility remaining with the authorised human decision-maker. Additional governance procedures are planned for implementation before operational deployment.

### Recommendations

It is recommended that the project maintain continuous engagement with the DFKI Ethics Team and the external ethics advisors throughout the development of the WP7 decision-support assistant. Regular monitoring of fairness, technical robustness, explainability, and privacy safeguards should remain part of the system lifecycle. As the system progresses towards higher technology readiness levels, any substantial changes to the system architecture, data sources, deployment environment, or intended use should include a new ethics evaluation.

## About the project

The InnovAite Slovakia project was launched in response to the call “Transformation and Innovation Consortia” announced by the Government Office of the Slovak Republic under the Recovery and Resilience Plan of the Slovak Republic (Component 9 – Investment 2: Supporting cooperation between companies, academia, and research and development organizations). The call aims to drive systemic transformation and increase the added value of key sectors of the Slovak economy through intensive collaboration among research institutions, innovative enterprises, the public sector, and internationally renowned partners.

InnovAite Slovakia seeks to build a dynamic and sustainable innovation ecosystem in the field of artificial intelligence (AI), effectively linking cutting-edge research with practical applications. The project emphasizes the development of AI solutions that are not only technologically advanced but also ethical, environmentally sustainable, and socially beneficial. The consortium brings together leading research centers, universities, and businesses from Slovakia, Germany, and the Czech Republic.

Key focus areas include the development of AI algorithms for improving building energy efficiency, enhancing traffic safety through video analytics, automating software development, driving digital transformation in the insurance sector, and validating functional prototypes. Special attention is also given to AI education, talent development, and the incubation of startups.

The project’s outcomes will significantly strengthen Slovakia’s position in research and innovation, reduce environmental burdens, increase the competitiveness of the national industry, and modernize public services. By bridging the public and private sectors with international research excellence, InnovAite Slovakia stands as a key instrument in addressing the social and economic challenges of today.



*“Funded by the EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project No. 09I02-03-V01-00029”*

Innovaite  
SLOVAKIA



[www.innovaite.sk](http://www.innovaite.sk)



Financované  
Európskou úniou  
NextGenerationEU

**PLÁN [OBNOVY]**



ÚRAD PODPRESEDU VLÁDY  
SLOVENSKEJ REPUBLIKY  
PRE PLÁN OBNOVY  
A ZNALOSTNÚ EKONOMIKU

  
**VAIA**  
VÝSKUMNÁ  
A INOVAČNÁ  
AUTORITA