AI for real-world network operation

WP4 – Validation and impact assessment

D4.1 – Evaluation and test protocols

AI4REALNET has received funding from European Union's Horizon Europe Research and Innovation programme under the Grant Agreement No 101119527, and from the Swiss State Secretariat for Education, Research and Innovation (SERI).

DOCUMENT INFORMATION

DOCUMENT	D4.1 – Evaluation and test protocol
TYPE	R — Document, report
DISTRIBUTION LEVEL	PU - Public
DUE DELIVERY DATE	31/03/2025
DATE OF DELIVERY	27/03/2025
VERSION	1.0
DELIVERABLE RESPONSIBLE	RTE
AUTHOR (S)	Bruno Lemetayer (RTE), Luca Saporetti (POLIMI), Manuel Schneider (FLATLAND), Ricardo Bessa (INESC TEC), Roman Liessner (DB)
OFFICIAL REVIEWER/s	Paolo Trucco (POLIMI), Jan Viebahn (TENNET), Sjoerd Kop (TENNET)

DOCUMENT HISTORY

VERSION	AUTHORS	DATE	CONTENT AND CHANGES
0.1	Bruno Lemetayer	29/08/2024	Creation of the document's structure
0.2	Bruno Lemetayer, Luca Saporetti, Manuel Schneider, Ricardo Bessa, Roman Liessner	28/02/2025	Assembling of each individual sections into the document
0.3	Bruno Lemetayer, Luca Saporetti, Manuel Schneider, Ricardo Bessa, Roman Liessner	10/03/2025	Release of the first version for internal review
0.4	Bruno Lemetayer	21/03/2025	Update following internal review
1.0	Ricardo Bessa	27/03/2025	Final version

ACKNOWLEDGEMENTS

NAME	PARTNER
Alberto Castagna	ENLITEAI
Anton Fuxjäger	ENLITEAI
Bruno Lemetayer	RTE
Cristina Felix	NAV
Christian Eichenberger	FLATLAND
Davide Lucioni	POLIMI
Eduardo Vilches	UKASSEL
Hélio Sales	NAV
Herke Van Hoof	UvA
Jan Viebahn	TENNET
Joaquim Geraldes	NAV
Joao Soares	NAV
Julia Usher	ZHAW
Kostiantyn Kucher	LiU
Kurt Brendlinger	FHG
Luca Saporetti	POLIMI
Manuel Meyer	FLATLAND
Manuel Schneider	FLATLAND
Milad Leyli-Abadi	IRTSX
Mohamed Hassouna	FHG & UKASSEL
Patrick Zinsli	FHNW
Timothy Tjhay	INESC TEC
Roman Liessner	DB
Samira Hamouche	FHNW
Sjoerd Kop	TENNET
Toni Waefler	FHNW
José Paulos	INESC TEC
Duarte Dias	INESC TEC

DISCLAIMER

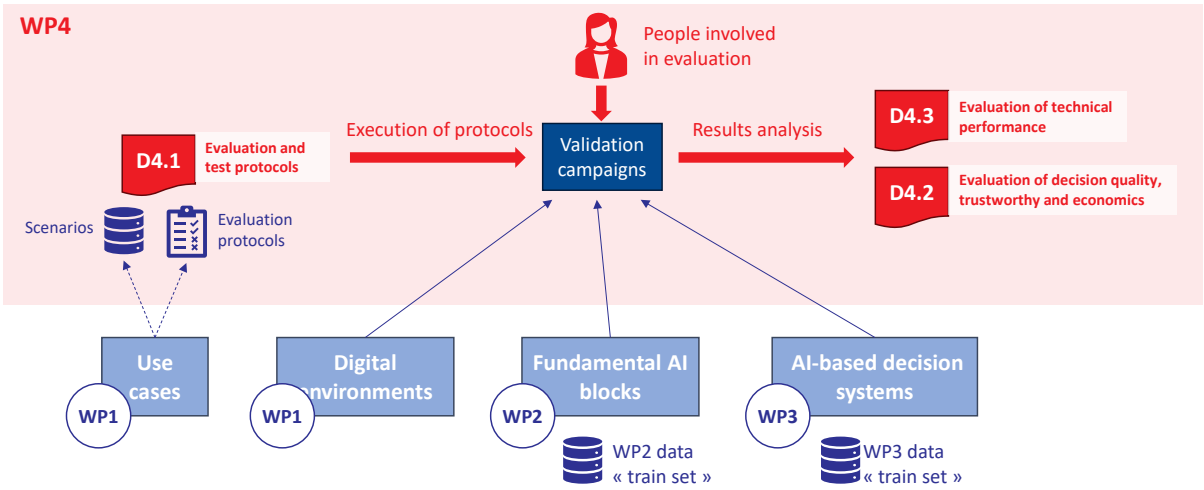
This project is funded by the European Union and SERI. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union and SERI. Neither the European Union nor the granting authority can be held responsible for them.

SUMMARY

Given the length of the document, a more detailed executive summary is provided below to offer a comprehensive overview of the key points, objectives, and outcomes for readers seeking a high-level understanding.

Within AI4REALNET project, the main goal of Work Package 4 (WP4) of AI4REALNET project is to evaluate the deliverables produced by the other Work Packages (see the figure below), namely:

- **Use Cases** and **Digital Environments** produced by Work Package 1 (WP1),
- **Fundamental AI blocks** produced by Work Package 2 (WP2),
- **AI-based decision systems** produced by Work Package 3 (WP3).



The whole WP4 evaluation work therefore mainly contributes to the main **AI4REALNET Objective O5** "Validate the proposed AI framework in a variety of use cases and realistic digital environments […]".

Deliverable D4.1 is the first deliverable produced by the WP4, and it has been written to define the methodology and organization of the evaluation carried out through two validation campaigns according to which all results will be produced, then reported in the two other deliverables from WP4 (D4.2, D4.3).

The document addresses **3 evaluation dimensions**, each linked to a relevant **main project Objective**:

- the qualitative and quantitative KPIs for the technical performance and scalability (Project Objective O2),
- the optimal balance between AI and human in the selected use cases (Project Objective O3),
- the robustness and safety of the AI solutions (Project Objective O4).

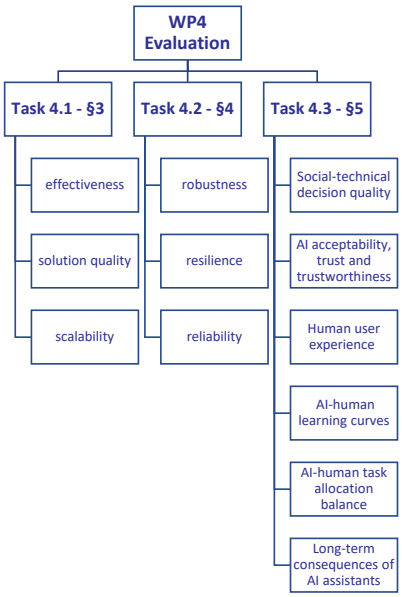
The writing of deliverable D4.1 has been distributed over the different tasks of WP4, that each cover different parts of the evaluation and thus contribute in a specific way to each of the 3 evaluation dimensions, as depicted in the figure below.

Evaluation **of technical performance** is covered by Task 4.1 and focuses on domain specific performance of infrastructure operation, associated to use case and project's objectives. The evaluation protocols in this task also evaluate the overall system efficiency of the combined human-AI team, and scalability of AI solutions during training and testing is also assessed.

Safety and robustness assessment is covered by Task 4.2 and uses a risk assessment performed on each of the project's domain. The evaluation protocols in this task address the robustness and adaptability of AI systems to domain shifts and perturbations. It measures the time required for AI to adapt to domain shifts and the performance loss due to these shifts, the accuracy of detecting out-of-domain operations and the generalization of policies to real-world scenarios, or sensitivity to changes in domain parameters and the performance drop after domain shifts. The stability of AI actions under perturbations and the robustness of explanations are also assessed.

Social-technical decision quality assessment is covered by Task 4.3, which focuses on human factors and the interaction between human operators and AI systems. The evaluation protocols in this task focus on human factors and the interaction between human operators and AI systems, by measuring the acceptance of the system by human users and the frequency of human interventions in AI decisions, or the satisfaction with decision support. Workload and cognitive performance of human operators are also evaluated, as well as the perceived impact on their skills and agency, and predicted long-term adoption of the AI assistant.

The resulting **evaluation objectives** are depicted in the figure below.

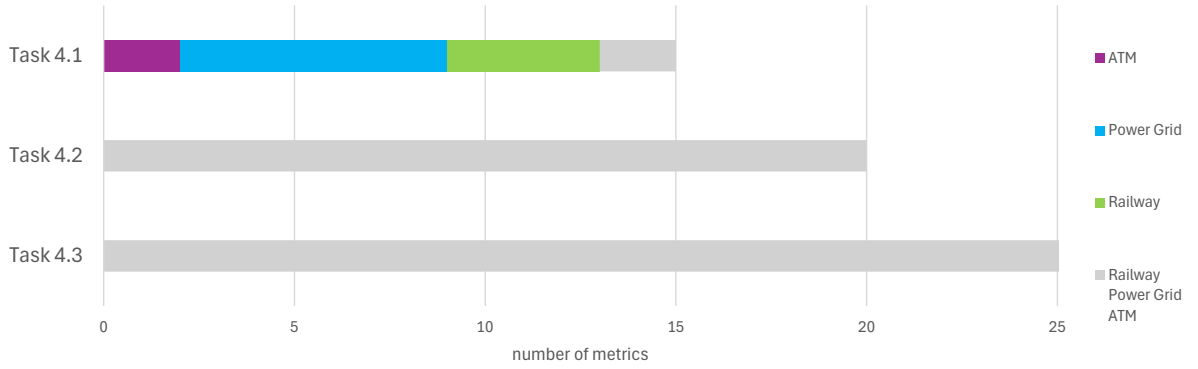


Economic and regulatory assessment principles are covered by Task 4.4, knowing that these will be further detailed once this task will be started.

EVALUATION PROTOCOLS

62 evaluation protocols covering 12 evaluation objectives and 3 project's objectives have been created (see the figure below) starting from previous work done in the deliverable D1.1 – Conceptual Framework and detailed by each Task. Each evaluation protocol has been associated with an **evaluation objective** to ensure that all intended project's objectives are covered.

The cross-domain opportunities have been used to define a very generic set of evaluation protocols, apart from the technical evaluation (Task 4.1) which is proper to each domain's specificities: all evaluation protocols for safety and robustness (Task 4.2) and social-technical decision quality (Task 4.3) are domain-generic (see the figure below).



Compared to the other types of evaluation, **social-technical decision quality evaluation** (carried out by Task 4.3) has the highest number of evaluation protocols, which shows the importance of this topic in the planned evaluation work. Especially, the high number of metrics associated to "human-user experience" and "social-technical decision quality" objectives will allow evaluating in detail **what AI can bring on top of current baseline to human operators**: "Perceived decision novelty evaluation metric" can for example show how AI can "augment" the human baseline, by bringing new ways of solving problems. At the same time, evaluation metrics like "Workload", "Assistant disturbance" or "Decision support satisfaction" will show if this isn't done to the detriment of the overall user experience. Socio-technical evaluations—such as perceived decision novelty, satisfaction with decision support, AI co-learning capability, human control and autonomy, and anticipated long-term adoption—will be linked to recent research on the benefits of AI augmentation. This includes studies on adaptive autonomous agents in human-AI teams and the impact of AI's adaptability and social role on professional efficacy and credit attribution, providing further evidence of AI co-learning's potential to enhance human operator performance.

TECHNICAL PERFORMANCE

Performance evaluation is focused on functional evaluation with a business and operation-oriented point of view. AI systems support real-time network operations by integrating information and forecasted conditions, enabling corrective and preventive actions at various automation levels. Manual actions are emphasized in the power grid domain, while higher automation levels are considered for railway and ATM domains. Each domain's network structure helps inform solution strategies and constraints. Performance evaluation is therefore defined specifically per domain and divided into **Solution quality** and **Effectiveness**.

Solution quality refers to the overall excellence and suitability of the solutions provided by the AI assistant. It involves several factors, including the accuracy of the AI's responses and recommendations, ensuring they are correct and reliable. Corresponding evaluation protocols are listed hereafter per domain:

Air Traffic Management

- KPI-RF-027: Reduction in delay,
- KPI-SS-032: System efficiency.

Power Grid

- KPI-AF-008: Assistant alert accuracy,
- KPI-NF-024: Network utilization,
- KPI-TS-035: Total decision time.

Railway

- KPI-DF-016: Delay reduction efficiency,
- KPI-PF-026: Punctuality,
- KPI-AF-029: AI Response time.

Effectiveness refers to the AI assistant's ability to achieve desired outcomes and meet its intended goals. This includes the efficiency with which the AI performs tasks, completing them quickly and with minimal resource consumption. Adaptability is another important factor, where the AI learns from interactions and improves over time, becoming more useful with continued use. Corresponding evaluation protocols are listed hereafter per domain:

Power Grid

- KPI-CF-012: Carbon intensity,
- KPI-TF-034: Topological action complexity,
- KPI-OF-036: Operation score,
- KPI-AS-068: Assistant adaptation to user preferences.

Railway

- KPI-NF-045: Network Impact Propagation.

Technical performance also includes measures of the **scalability** of an AI assistant: this involves evaluating its ability to handle increasingly complex scenarios without compromising performance. Two main aspects of scalability are considered, and defined in the same way for all domains, taking advantage of the shared notion of “infrastructure”, which can always be associated with a sizing metric:

- **KPI-AF-050**: Scalability at training time, which measures how much more time and resources are needed to train a model on larger scenarios,
- **KPI-AF-051**: Scalability at testing time, which evaluates the ability of the agent to handle larger scenarios while evaluating potential additional time costs for the decision making.

SAFETY AND ROBUSTNESS

Given the criticality of power grids, railway networks, and air traffic management infrastructures, ensuring that AI models function accurately under various conditions, including adversarial perturbations and environmental changes, is paramount. This is ensured by assessing AI-based systems’ **robustness**, **resilience**, and **reliability**.

The applied methodology follows a structured approach based on **standardisation frameworks** such as ISO/IEC 24029-2 and the AI Act requirements. The first step involves **risk identification and**

assessment, where various technical risks are identified, including data drift, adversarial attacks, and reliability concerns. Vulnerabilities in different AI components, ranging from model inputs to decision outputs, are systematically evaluated.

This risk analysis reveals that a common aspect of these three infrastructures is that the critical technical risk source is **perturbations in the state/input space**. These perturbations are more frequent (e.g., information collected from different internal and external data sources) and have a higher potential impact, as they cannot be mitigated via model replacement or retraining. Another risk mentioned in the power grid – and that leads to one specific use case – but that could be applied to other infrastructures is out-of-domain data, particularly when digital environments are used for the initial training of the AI-based decision system. Significant differences might occur between digital and real environments.

Once the risks are identified, **adversarial datasets** are generated to simulate cyberattacks, sensor failures, and environmental disruptions. This involves using perturbation agents that create controlled variations in AI inputs, allowing for a thorough examination of system performance under stress. Domain-specific perturbation models are developed for power grids, railways, and air traffic control to ensure that AI models remain functional despite real-world uncertainties.

The **perturbation agents** considered in this evaluation include a gradient estimation perturbation agent, which estimates gradients to craft adversarial inputs that minimise the likelihood of correct AI decisions. A reinforcement learning-based perturbation agent is also employed, training an RL model to apply minimal but impactful perturbations that disrupt AI model performance. An action perturbing agent modifies the actions suggested by AI to simulate human intervention or execution deviations, while a communication perturbing agent introduces errors or delays in inter-agent communication, reflecting network or protocol issues. Additionally, domain-specific perturbation agents tailor perturbation mechanisms to each application domain. For power grids, simulated noisy measurements and state estimation errors are introduced. In railway networks, perturbations include track availability and occupancy misreports, train positioning errors, and schedule disruptions. In air traffic management, adverse weather conditions, flight entry delays, and sectorization disruptions are considered.

Then multiple evaluation metrics are defined to measure robustness, resilience, and reliability under controlled perturbations and comparing it with baseline operational scenarios.

Robustness evaluations provide critical insights into how AI models respond to stress conditions and include performance degradation under adversarial conditions, stability of decisions under input perturbations, and the rate of successful adversarial attacks. The specific metrics considered include:

- KPI-DT-069: Performance degradation under adversarial conditions,
- KPI-RT-058: Impact of partial human intervention on AI decisions,
- KPI-FT-070: Stability of AI outputs under input perturbations,
- KPI-ST-071: Similarity scores for modified AI decisions,
- KPI-VT-073: Proportion of adversarial examples causing action changes.

Resilience metrics ensure that AI systems can recover from adverse events effectively and thus focus on the area between performance degradation and recovery curves, and maximum deviation from nominal performance. The specific metrics considered include:

- KPI-AT-074: Area between the reward curves of unperturbed and perturbed AI systems,
- KPI-DT-075: Duration of degradation and recovery stages,
- KPI-RT-076: Minimum and maximum rewards during degradation and restoration,
- KPI-ST-077: Similarity of the operational environment state over time.

Reliability evaluations focus on the long-term sustainability and operational consistency of AI models in real-world applications, by assessing mean time between failures (MTBF), accuracy detecting out-of-distribution (OOD) data, and adaptation time to environmental shifts. The specific metrics considered include:

- KPI-DT-057: Accuracy in detecting domain shifts and out-of-distribution data,
- KPI-DT-052: Adaptation time required to recover from domain shifts,
- KPI-DT-056: Sensitivity of AI performance to changes in domain parameters,
- KPI-DT-055: Robustness of AI policies under domain shifts,
- KPI-DT-053: Generalization gap between training and test domains,
- KPI-DF-090: Domain shift forgetting rate.

Collectively, these evaluations contribute to developing AI solutions that are technically robust, resilient to external disruptions, and capable of maintaining performance integrity over extended periods.

AI systems should be subjected to continuous adversarial testing to identify potential vulnerabilities before deployment. By incorporating adversarial dataset generation as a core component of the evaluation process, developers can preemptively address issues that may arise in real-world applications.

SOCIAL-TECHNICAL DECISION QUALITY

While ensuring high levels of performance from a technical standpoint is a crucial foundation for AI solutions, it is equally important to consider the intended context of real-world applications of such AI solutions to be developed within the scope of the AI4REALNET project.

All evaluations have been grouped according to different objectives: Social-technical decision quality, AI acceptability, trust, and trustworthiness, Human-user experience, AI and human learning curves, Task allocation balance, and Long-term consequences of AI-assistants.

Social-technical decision quality objective puts the stress on *decision quality* within the context of human operator interaction with the AI assistant (cf. “The efficiency of combined human-AI performance” and “Quality of AI-based solutions perceived by human operators” in Annex 4 of D1.1). The evaluation protocols in this cluster represent the number of human operator interventions and scale of necessary revisions for generated decisions. The specific metrics considered include:

- KPI-HS-003: Human intervention frequency

- KPI-SS-030: Significance of human revisions
- KPI-PS-089: Perceived decision novelty

AI acceptability, trust, and trustworthiness evaluation objective focuses on human operator-oriented indicators of trust and acceptability for the AI assistant in general as well agreement with and trustworthiness of specific AI-generated decisions. As part of this objective, the following concerns are addressed:

- Trust and acceptability for the AI assistant in general,
- Agreement with AI decisions,
- Trust for individual AI decisions,
- Perceived decision explainability.

The specific metrics considered include:

- KPI-AS-002: Acceptance,
- KPI-TS-039: Trust towards the AI tool,
- KPI-AS-005: Agreement score,
- KPI-TS-038: Trust in AI solutions score,
- KPI-CS-013: Comprehensibility.

Human-user experience evaluation objective focuses on human operator-oriented indicators of user experience, including concerns such as workload and stress; alignment with and support for the operators' cognitive processes; and user motivation and satisfaction. The respective evaluation metrics are measured with quantitative psychophysiological indicators and qualitative assessments.

As part of this objective, the following concerns are addressed:

- Workload and stress,
- Alignment and support for the operators' cognitive processes,
- User motivation and satisfaction.

The specific metrics considered include:

- KPI-WS-040: Workload,
- KPI-AS-009: Assistant disturbance,
- KPI-CS-049: Cognitive performance and stress,
- KPI-AS-001: Ability to anticipate,
- KPI-SS-031: Situation awareness,
- KPI-HS-023: Human response time,
- KPI-HS-022: Human motivation,
- KPI-DS-015: Decision support satisfaction.

AI and human learning curves evaluation objective focuses on several aspects of the AI and human operator learning:

- AI co-learning capability, AI ability to adapt to the operators’ preferences,

- Human learning, human operators' perceived learning opportunities when working with AI assistant.

The specific metrics considered include:

- KPI-AS-006: AI co-learning capability,
- KPI-HS-021: Human learning.

Task allocation balance evaluation objective focuses on the optimal balance between AI and human, and requirements in terms of new task allocation. The specific metrics considered include:

- KPI-HS-018: Human control/autonomy over the process,
- KPI-IS-041: Impact on workload.

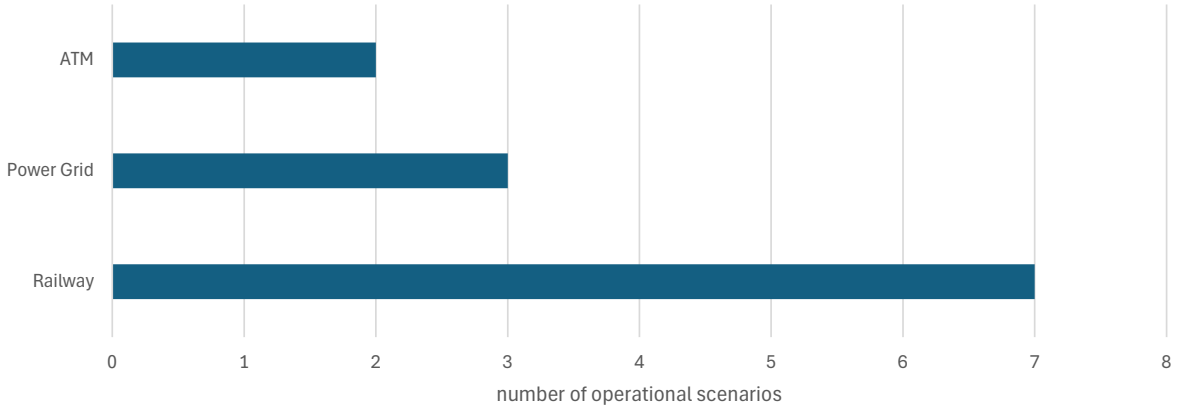
Long-term consequences of AI-assistants evaluation objective focuses on perceived and predicted long-term consequences of AI assistant adoption. Due to the potential variation in availability of user study participants and stakeholders, the emphasis here is made on evaluations that can be done within the scope of individual studies rather than necessarily over a series of experiments; the evaluations focusing on reflections over the ongoing/past deployments and pilot studies would thus benefit from being assessed closer to the final rather than initial stages of the project. The specific metrics considered include:

- KPI-RS-091: Reflection on operator trust,
- KPI-RS-092: Reflection on operator agency,
- KPI-RS-093: Reflection on operator de-skilling,
- KPI-RS-094: Reflection on over-reliance,
- KPI-RS-095: Reflection on additional training,
- KPI-RS-096: Reflection on biases,

KPI-PS-097: Predicted long-term adoption.

OPERATIONAL SCENARIOS

Starting from use cases defined in the deliverable D1.1 – Conceptual Framework, 12 realistic and representative operational testing scenarios have been defined for each domain (see figure below) and will be used throughout the validation campaigns.



These scenarios have been defined to ensure that all objectives of deliverable D4.1 – Evaluation and test protocols are covered (see table below): Technical performance and scalability, Robustness and safety of the AI solutions, Optimal balance between AI and human.

D4.1 objective	Scenarios
Technical performance and scalability	**ATM**: Military Reservation OPSCE-UC2.A-1-011 **Power Grid**: Remedial actions OPSCE-UC1.P-1-001 **Railway**: Reactive Re-Scheduling OPSCE-UC2.R-1-007, Proactive re-scheduling OPSCE-UC2.R-3-009
Robustness and safety of the AI solutions	**ATM**: Adverse weather conditions OPSCE-UC1.A-1-012 **Power Grid**: Real world conditions OPSCE-UC2.P-1-003 **Railway**: Re-Scheduling at the occurrence of infrastructure malfunction OPSCE-UC1.R-1-004, Emergency response to adverse weather conditions OPSCE-UC1.R-2-005, Closure of a large station OPSCE-UC1.R-3-006
Optimal balance between AI and human	**ATM**: Military Reservation OPSCE-UC2.A-1-011, Adverse weather conditions OPSCE-UC1.A-1-012 **Power Grid**: AI assistant learns OPSCE-UC1.P-2-002 **Railway**: Co-learning for reactive re-scheduling OPSCE-UC2.R-2-008, Co-learning for proactive re-scheduling OPSCE-UC2.R-4-010

Each scenario is instantiated using one or more datasets, that are either built within the range of ordinary condition (i.e. within the distribution of input space data) or are specifically designed to test the limits of the systems (thus outside or at the extremity of distribution of input space data).

The possibility to instantiate digital environments by using real data to simulate realistic conditions is different per domain. While ATM and Railway domains can easily reuse real open-sourced data, using real data from Power Grid is more difficult, especially due to confidentiality issues for generation units: depending on legal constraints, it is therefore envisaged that real data could be used as confidential data that won't be open sourced.

ECONOMIC ASSESSEMENT

ABC methodology is a method used to analyze and model the operational processes of a system by linking the activities performed with the necessary resources and associated costs. It will be applied to assess the impact of AI solutions in critical contexts of the three selected domains. The main steps of the methodology to be applied are the following:

1. **Analyze processes and activities**: Understand the key processes of each application domain and decompose them into individual activities.
2. **Link activities to resources**: Gather key execution information for each activity (e.g., allocated resources, timelines, interdependencies with other activities) and estimate the costs associated with resources and activities.
3. **Quantify economic impacts**: Evaluate the efficiency of activities and the impact of introducing new AI solutions and calculate economic indicators such as the total cost of the solution, return on investment (ROI), and the added value of changes (qualitative and quantitative).
4. **Simulate scenarios**: Depending on the data available, sensitivity analysis may be implemented to evaluate the economic outcomes of technology adoption under different scenarios (Worst, Best, Base scenarios).

REGULATORY ASSESSMENT

The **regulatory assessment** will be mainly based on current applicable EU regulations and EU Agency for Cybersecurity recommendations, and will consider:

- Issues related to ethics, data protection, dataset, algorithm bias, infrastructure vulnerabilities, etc.
- Impacts on the society, workers, individuals, minorities, etc., focusing on any qualms or worries due to the adoption of AI, and relying on stakeholders' consultation,
- Potential evolution of employment law and workers' rights with AI,
- Other existing ethical, legal, social and impact research methodologies, such as ELSA or Responsible Research and Innovation (RRI) methodology,
- Relevant work from sister projects (same EU funding programme).

The assessment will be risk based, i.e. impact, consequences and likelihood will be estimated based on experts' opinions (legal and AI experts), considering the innovation of the project, and providing practical experiences in AI model development.

Following the regulatory assessment, a **mitigation plan** will be proposed. This plan will include:

- what is already implemented,
- what is planned to be implemented and the deadline,
- what is not implemented nor planned,
- mitigation actions.

The mitigation plan will be discussed with all relevant stakeholders, from people involved in the design of algorithms and interfaces, to the operational experts considering the use of AI systems.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS AND ACRONYMS

Acronym	Definition
AI	Artificial Intelligence
AI HELG	High-level Expert Group on Artificial Intelligence
ALTAI	Assessment List for Trustworthy Artificial Intelligence
ANN	Artificial Neural Networks
ANSP	Air Navigation Service Provider
ATC	Air Traffic Control
ATCO	Tactical Air Traffic Controller
ATM	Air Traffic Management
AUGT	Automated Urban-Guided Transport
CAB	Cockpit and Bidirectional Assistant
CSE	Cognitive Systems Engineering
EID	Ecological Interface Design
ENTSO-E	European Network of Transmission System Operators for Electricity
FIR	Flight Information Region
FMP	Flow Management Position
GDPR	General Data Protection Regulation
GoA	Grade of Automation
H-AI	Human-AI
H-H	Human-human
ICAO	Internation Civil Aviation Organization
IEC	International Electrotechnical Commission
IEEE	Institute of Electrical and Electronics Engineers
ISO	International Organization for Standardization
JCF	Joint Control Framework
KPI	Key Performance Indicator
LACC	Level of Autonomy in Cognitive Control
LOA	Level of Automation
MARL	Multi-agent Reinforcement Learning
MDP	Markov Decision Process
ML	Machine Learning
OoS	Out-of-Scope
OPF	Optimal Power Flow
POC	Proof of Concept
RAMS	Reliability, Availability, Maintainability, and Safety
RL	Reinforcement Learning
RUOM	Railway Undertaking Operating Manager
SAGAT	Situation Awareness Global Assessment Technique
SCADA	Supervisory Control and Data Acquisition
SL	Supervised Learning
TAI	Trustworthy AI
TEF	Testing and Experimentation Facilities
TSO	Transmission System Operator
UC	Use Case
UQ	Uncertainty Quantification
XAI	Explainable AI

1. INTRODUCTION

Within AI4REALNET project, the main goal of Work Package 4 (WP4) is to evaluate the deliverables produced by the other Work Packages (see Figure 1 – contribution of WP4 in AI4REALNET project), namely:

- Use Cases and Digital Environments produced by Work Package 1 (WP1),
- Fundamental AI blocks produced by Work Package 2 (WP2),
- AI-based decision systems produced by Work Package 3 (WP3).

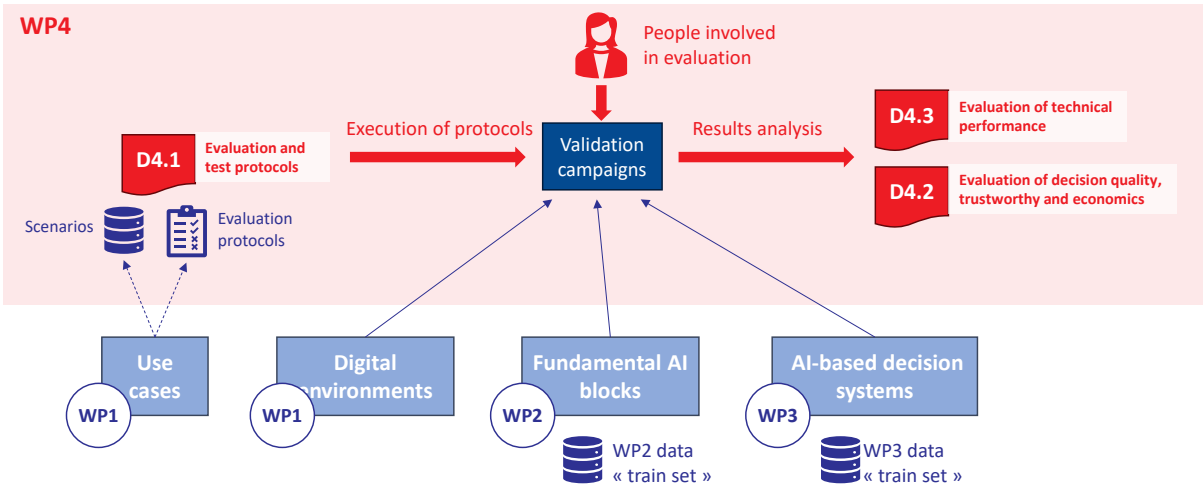


FIGURE 1 – CONTRIBUTION OF WP4 IN AI4REALNET PROJECT

According to the Conceptual Framework of AI4REALNET project defined in deliverable D1.1, evaluating an AI assistant in critical environments requires a holistic approach that considers technical, social, ethical, economic and regulatory aspects, to ensure that AI improves the efficiency and safety of operations while maintaining the trust and control of human operators.

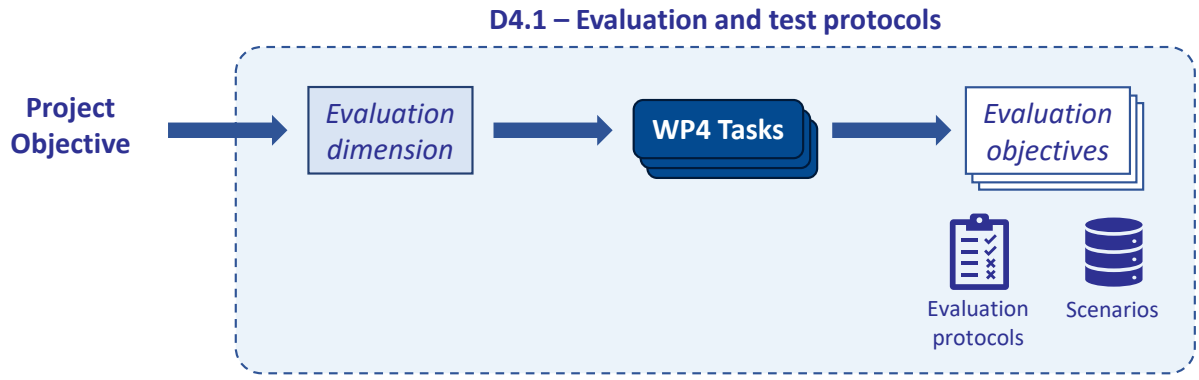


FIGURE 2 – OVERALL D4.1 APPROACH

Consequently, this document (D4.1) has been written to define the methodology and organization of the evaluation carried out through two validation campaigns. It contains (i) an overall methodological approach, with overall principles and organization of evaluation work, technical setup, execution and

reporting. It defines (ii) specific protocols to address the different evaluation objectives and peculiarities of the use cases: these protocols will be executed according to (i). Corresponding results will be produced, then reported in the two other deliverables from WP4 (D4.2, D4.3).

The overall approach to write this document (see Figure 2) started with considering how it will contribute to the main **AI4REALNET project Objective**s (O1 to O5), and how each main AI4REALNET project Objective can be translated into a relevant **evaluation dimension** (see Table 1).

Project Objective	Description	Evaluation dimension	WP4 contribution
O1	Design and develop a domain-general AI framework, within the formalism of sequential decision making, with special focus to human-AI interaction (human-assistant, co-learning, full AI-based control)	none[1]	none[2]
O2	Deliver novel variants of SL/RL for large-scale complex networks, exploiting domain knowledge, hierarchical and distributed problem decomposition to increase scalability in realistic networks, and supported by a set of functions and HMI for explainable, algorithmically transparent, and interpretable RL	Assess qualitative and quantitative KPIs for the technical performance and scalability	**Task 4.1**: evaluate AI technical performance, including scalability **Task 4.3**: evaluate trustworthiness
O3	Develop human-assistance and co-learning strategies between humans and AI that augment decision making capabilities under risk and uncertainty	Assess optimal balance between AI and human in the selected use cases	**Task 4.3**: evaluate social-technical decision quality **Task 4.4**: economic assessment
O4	Develop autonomous AI agents with safety and robustness as primary requirements	Assess robustness and safety of the AI solutions	**Task 4.2**: evaluate safety and robustness of AI solutions **Task 4.4**: regulatory assessment

[1] This objective corresponds to WP1, Task 1.1

[2] This objective corresponds to WP1, Task 1.1

Project Objective	Description	Evaluation dimension	WP4 contribution
O5	Validate the proposed AI framework in a variety of use cases and realistic digital environments for operation of critical infrastructures, along different dimensions (technical, SSH, economic and regulation), and increase social, academic (AI community) and business awareness to AI potential	All	**Tasks 4.1**, **4.2**, **4.3** and **4.4** evaluations

TABLE 1 – CONTRIBUTION OF WP4 TO PROJECT'S OBJECTIVES

AI4REALNET Objective O1 "Design and develop a domain-general AI framework […]" is specifically covered by WP1, Task 1.1. The WP4 evaluation work contributes as a whole to the AI4REALNET Objective O5 "Validate the proposed AI framework in a variety of use cases and realistic digital environments […]". D4.1 therefore specifically addresses **3 main evaluation dimensions** and their corresponding **AI4REALNET project Objective**s that are described hereafter.

Assess qualitative and quantitative KPIs for the technical performance and scalability *(Objective O2 "Deliver novel variants of SL/RL for large-scale complex networks […]")*

The evaluation must focus on the system's performance in both normal and abnormal conditions, to assess its suitability for use cases' objectives. In case of system failure, the **effectiveness** of the AI and the added value of knowledge information (**quality** of the solution) must be ensured. An AI system is considered reliable if it behaves as expected, even for novel inputs on which it has not been trained or tested. AI systems must be capable of handling large and realistic scenarios: training and inference methods, as well as algorithms, thus must consider **scalability** constraints.

Assess optimal balance between AI and human in the selected use cases *(Objective O3 "Develop human-assistance and co-learning strategies between humans and AI that augment decision making capabilities […]")*

AI systems must be designed for seamless **interaction with human operators**, respecting their autonomy and fostering appropriate trust. This includes various modes of interaction, ranging from full human control to full AI-based control, including co-learning between AI and humans. The goal is for AI to assist, but not replace, operators. It is essential for humans to understand the reasoning behind AI decisions, especially in critical infrastructures. **Explainability** must be evaluated in terms of fidelity (accuracy of the explanation compared to the decision-making process), intelligibility, and actionability. Explanations must be provided in a timely manner, i.e., without delaying the decision-making process. Operators should not be **demotivated** or **deskilled** due to interaction with AI. Evaluations should also measure the relevance of AI recommendations, the impact on the operator's **workload**, and the level of **trust**.

Besides autonomy, workload, trust, and human learning, the beneficial interaction for AI and human is also measured in terms of long-term **economic benefits**, through reduction of operational costs, improvement of operational efficiency, reduction of human workload or optimization of resource

utilization. AI should effectively assist humans without compromising their skills or creating excessive dependence.

Assess robustness and safety of the AI solutions *(Objective O4 "Develop autonomous AI agents with safety and robustness as primary requirements")*

The AI system's ability to maintain its performance level in the face of natural or adversarial disturbances is paramount. Technical **robustness** includes the capacity to manage bad or missing data, model perturbations, or adversarial attacks. It can be evaluated by measuring the variance of the output or the KPI in response to perturbations, or through an adversarial agent. AI must be designed not to increase **safety** risks, whether physical or cyber. The system's resilience to attacks and data **security** must be assessed. It is important to comply with **regulatory** requirements, such as the **European AI Act** and domain-specific safety standards. Quantifying and managing the **uncertainties** inherent in AI models and real-world data is necessary to help in decision-making in uncertain contexts. This includes epistemic uncertainty (related to the model) and aleatoric uncertainty (related to the data).

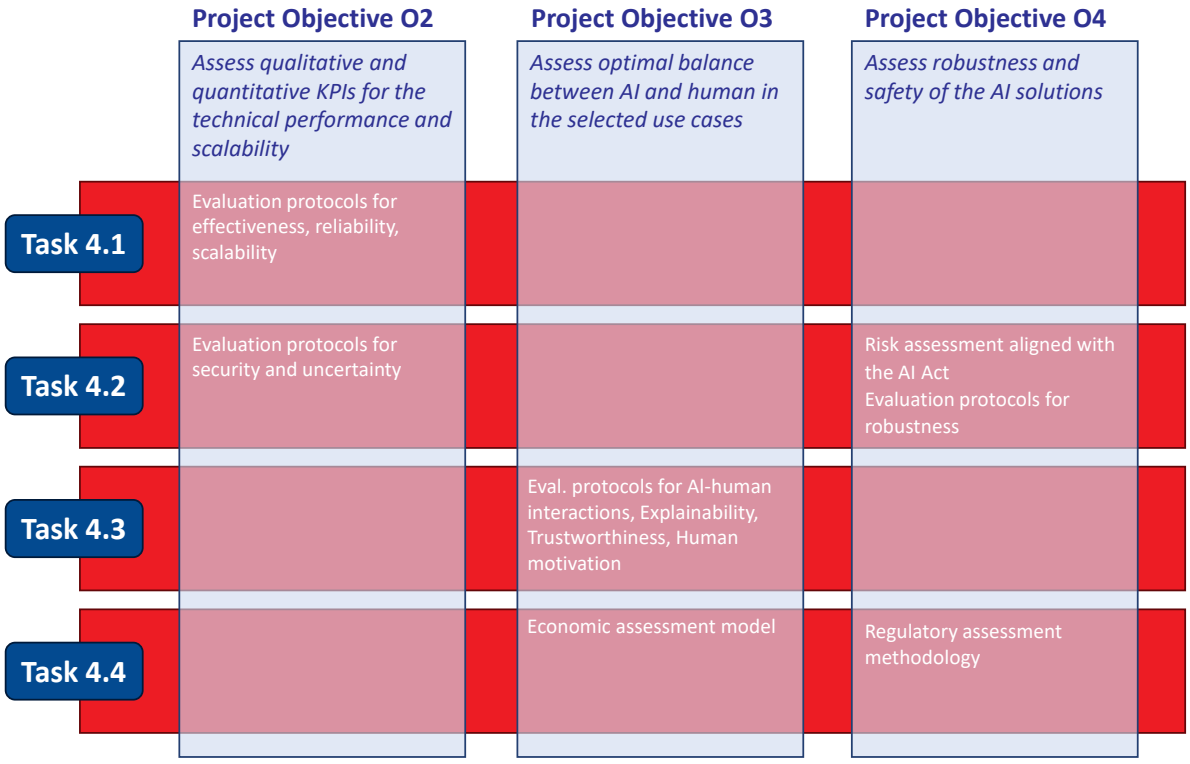


FIGURE 3 – CONTRIBUTION OF EACH TASK TO EVALUATION DIMENSIONS

The writing of D4.1 has been distributed over the different tasks of WP4, that each cover different parts of the evaluation and thus contribute in a specific way to each of the 3 evaluation dimensions, as outlined in Table 1 and depicted in Figure 3. Each task has detailed the relevant evaluation dimensions into detailed evaluation objectives, by describing associated **evaluation protocols** and **scenarios**.

The different sections of the document have been structured accordingly and are organized as follows (see Figure 4):

- Section §2 “Validation plan” defines overall principles and organization of evaluation work,
- Section §3 “Evaluation of technical performance” details contribution of Task 4.1 to the evaluation, using a common template to define evaluation protocols, and a common template to define scenarios,
- Section §4 “Safety and robustness assessment” details contribution of Task 4.2 to the evaluation, using a common template to perform risk assessment and to define evaluation protocols,
- Section §5 “Social-technical decision quality” details contribution of Task 4.3 to the evaluation, using a common template to define evaluation protocols,
- Section §6 “Economic and regulatory assessment” details contribution of Task 4.4 to the evaluation,
- Section §7 “Execution and reporting” defines how evaluations are technically setup, run and reported. Furthermore, it defines the project responsibilities during execution.

Annexes contain the details of all evaluation protocols, scenarios and assessment used and referenced in sections §3, §4 and §5 (see Figure 4).

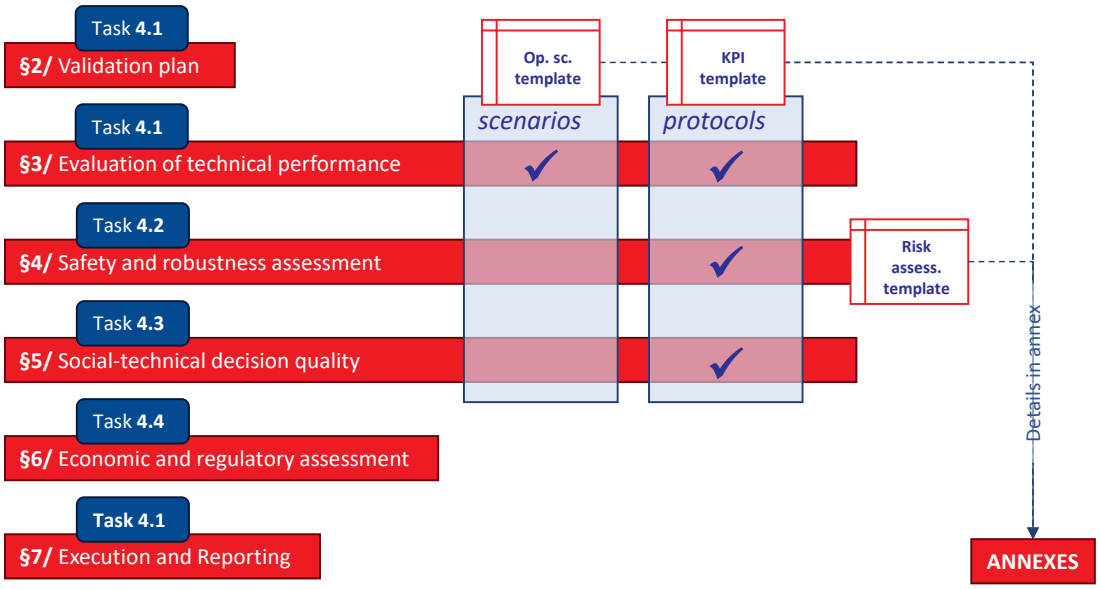


FIGURE 4 – D4.1 DOCUMENT’S STRUCTURE

2. VALIDATION PLAN

This section defines overall principles and organization of evaluation work. It does not apply to economic and regulatory assessment which is carried out outside the technical setup (see §6).

It presents a systematic approach to evaluate AI-driven solutions across multiple domains – Air Traffic Management, Power Grid and Railway – designed to assess technical performance, operational efficiency, and the interaction between human operators and AI systems. This approach combines digital simulation, live operational data, and comprehensive testing methodologies to ensure that AI integration meets stringent regulatory requirements and industry standards.

The validation process emphasizes key performance objectives, including safety and reliability, robustness under varied conditions, scalability across complex systems, and optimized human–machine collaboration.

Next subsections describe the overall validation strategy, testing methodologies, and a two-phase validation campaign which ensures a thorough evaluation of the AI systems, providing actionable insights and recommendations for full-scale deployment.

2.1 VALIDATION FRAMEWORK

The validation framework applied to AI-driven solutions produced by WP2 and WP3 (see Figure 5 – Validation framework) is designed to provide a holistic assessment across diverse operational scenarios. By using **simulated environments** in the **validation campaign hub** (see §7.3 - Technical setup) and **operational scenarios** to execute **detailed evaluation protocols** (see §7.2 - Procedure), the validation framework integrates a variety of data sources and testing methodologies to provide a comprehensive evaluation of AI performance across multiple domains, ensuring that every aspect of the AI solution is rigorously examined and that the **evaluation objectives** are covered.

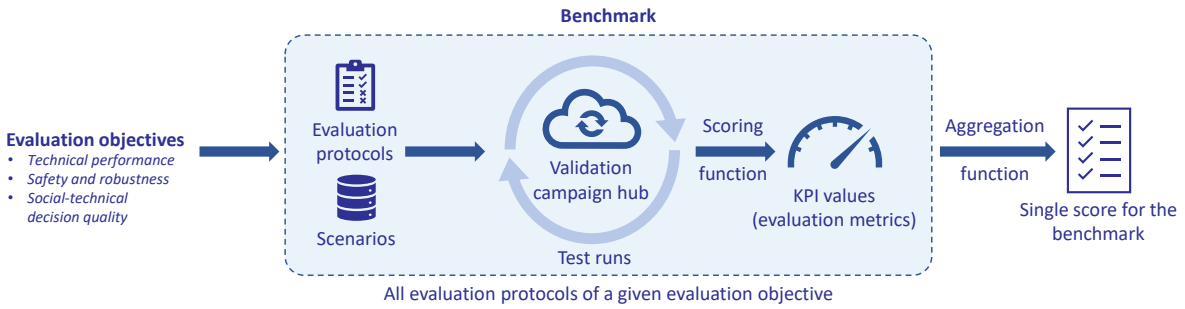


FIGURE 5 – VALIDATION FRAMEWORK

Starting from KPIs defined for use cases in deliverable D1.1 (AI4REALNET framework and use cases), detailed **evaluation protocols** have been developed, with accompanying documents outlining the description, calculation, and application steps for a wide array of performance variables and human evaluation (e.g. questionnaire), which include:

- Accuracy & Precision: Evaluating the correctness of AI-generated outputs in various decision-making contexts,
- Robustness & Resilience: Assessing the system's ability to maintain functionality amid unexpected disruptions or environmental changes.
- Explainability & Transparency: Ensuring that the decision-making process is clear and that outputs are interpretable by end-users.
- Latency & Response Time: Measuring the system's speed in processing data and reacting to dynamic conditions.
- Scalability: Evaluating the system's performance as operational complexity and data volumes increase.
- User Trust & Usability: Capturing the satisfaction and confidence of end-users through structured feedback sessions.

These protocols and their associated objectives are further detailed in sections §2 to §5, and yield **evaluation metrics** as **Key Performance Indicators (KPIs)** which are integral to monitoring system performance over time. Each evaluation protocol document (see Annex 1) is given a unique identifier and contains a detailed description and calculation methodology. This approach ensures that a wide range of variables is covered, and that performance metrics are consistently monitored throughout the validation process.

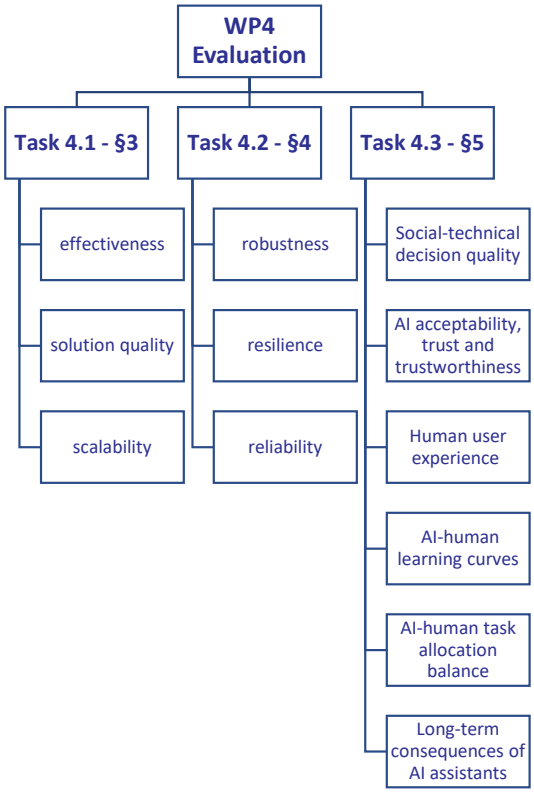


FIGURE 6 – WP4 EVALUATION OBJECTIVES

To ensure a consistent scope, each evaluation protocol has been associated with a specific evaluation objective (see Figure 6 – WP4 Evaluation objectives), itself corresponding to a main project objective (see Figure 2).

Moreover, detailed documents describing the **operational scenarios used in the validation tests** have been defined (see Annex 2). These documents outline the conditions, parameters, and expected outcomes for each test scenario, ensuring that tests are conducted consistently and that results are reproducible across different operational fields. They define the type of data used, either real-world operational data (historical data), or synthetic data generated for the project.

The validation framework is inherently **domain-agnostic** and is designed to be applied uniformly across Air Traffic Management, Power Grid and Railway domains, apart from domain-specific evaluation linked to technical performances. Although each domain presents unique operational challenges, the underlying principles of safety, efficiency, and reliability remain consistent. This approach ensures that the validation methodologies are robust, adaptable, and scalable.

All validation activities **comply with prevailing regulatory frameworks and industry standards**. The process is designed to comply with guidelines such as the EU AI Act and relevant ISO AI standards. This adherence guarantees that every stage of the validation process—from preliminary testing to final deployment—meets high safety and regulatory benchmarks.

2.2 TESTING AI ASSISTANTS

The evaluation of individual AI agents is performed with different configurations.

In **isolated module testing,** each AI module undergoes isolated testing to verify its standalone functionality. This process includes comprehensive unit and integration tests designed to identify and resolve issues at the module level. Isolated testing is essential for debugging and optimizing individual components before they are integrated into the larger system. This configuration is applied before the evaluation work of WP4.

Following isolated testing, in **scenario-based validation** (see §3 - Technical performance and scalability) AI agents are subjected to a series of simulated scenarios that mimic real-world operational conditions. These scenarios include both routine and edge-case situations—such as unexpected system failures, data anomalies, and extreme environmental changes. By exposing the AI models to a broad range of scenarios, the validation process ensures that the systems are robust and adaptable.

Adversarial and stress testing (see §4 - Safety and robustness) evaluates the robustness of AI agents through adversarial testing, where the systems are deliberately challenged with disruptive inputs. This stress testing is crucial for identifying potential vulnerabilities and ensuring that the AI can maintain operational integrity even under adverse conditions.

Human–AI collaboration evaluation (see §5 - Social-technical decision quality) assesses the interaction between human operators and AI systems is a key component of the validation process. Structured evaluation sessions are conducted in which operators engage with the AI in realistic scenarios. These sessions measure the clarity and effectiveness of AI outputs, the ease of use of the

interface, and the overall impact on decision-making. Feedback is collected systematically to inform further improvements in both the AI algorithms and the system interface.

2.3 VALIDATION CAMPAIGNS

A core component of the validation process is the continuous **iterative improvement and feedback loops** from simulation tests and real-world user interactions. Regular feedback loops with domain experts and technical teams ensure that the AI systems are iteratively refined. This iterative approach enables ongoing improvements in algorithm performance, system interface design, and overall operational integration. The validation process is executed with an **incremental approach** in **two validation campaigns** (see Figure 7 – Validation campaigns), which minimizes risk and provides a structured pathway for continuous improvement. Each phase builds on the insights and performance metrics gathered from the previous one.

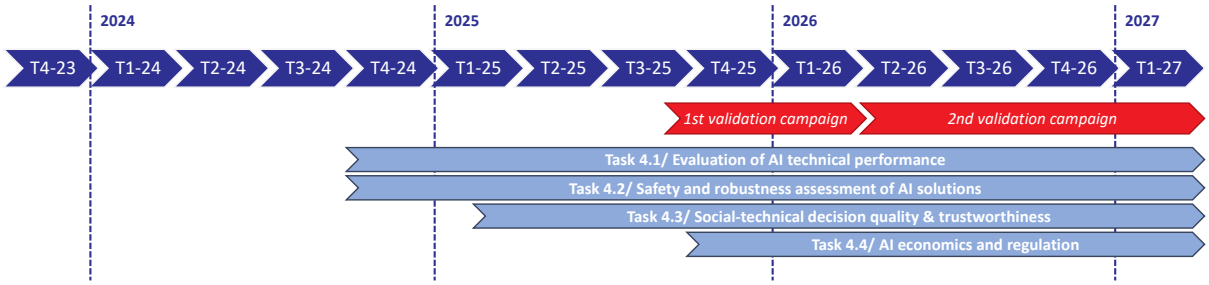


FIGURE 7 – VALIDATION CAMPAIGNS

2.3.1 1ST VALIDATION CAMPAIGN

The first validation campaign (M24-M30) serves as the initial phase for testing the feasibility and preliminary performance of the AI-driven solutions. It assesses early results of the project and serves as a compass for the remainder of the project. While only a subset of the evaluations will be performed during the first campaign, the scoring approach outlined in §7.4 will provide indications which areas show already good results, which areas fulfil the acceptance criteria but can be improved, and which areas will need special attention until the second validation campaign.

The test specifications, evaluation protocols, and similar aspects will be refined based on the learnings of the first validation campaign. Conducted in controlled, simulated environments, this phase will establish baseline performance metrics and gathering early feedback from stakeholders.

2.3.1.1 GOALS AND SCOPE

The primary objectives of the first validation campaign are following:

- **Assess feasibility** and confirm the fundamental functionality and decision-making capabilities of the AI models under controlled conditions,
- **Simulate environment testing** and use high-fidelity digital simulations to replicate operational scenarios across the three application domains,

- **Conduct training sessions** to familiarize end-users with the AI interface and operational protocols.

2.3.1.2 EXPECTED OUTCOMES

The anticipated outcomes from the first validation campaign include:

- **Identification of model limitations** and areas where the AI requires refinement,
- **User feedback for improvement** in both AI algorithms and user interfaces,
- **Establishment of baseline performance** with creation of a set of baselines, which will serve as benchmarks for measuring progress in the second validation phase,
- **Preparation of the system** for full-scale testing in live operational settings.

2.3.2 2ND VALIDATION CAMPAIGN

The second validation campaign (M30-M42) leverages the foundation built during the first campaign and focuses on incremental improvements, deeper complexity, and a more rigorous comparative analysis. It serves as the final assessment of the outcomes of the AI4REALNET project. The results of the first validation campaign will be included in the second campaign and can be improved during the second validation campaign. In addition, the evaluation protocols that were not addressed in the first validation campaign will be included in the second validation campaign. Like in the first validation campaign, tests can be run multiple times. Unlike the first validation campaign, the scores reached at the end of the second campaign will be used for the overall assessment and represent a final quality score.

This phase is designed to ensure that the AI systems not only meet initial safety and performance criteria but also demonstrate measurable progress in robustness, scalability, explainability, and Human-AI collaboration.

2.3.2.1 DIFFERENCES WITH 1ST VALIDATION CAMPAIGN

The second campaign introduces **more complex and diverse scenarios**, testing **scalability and robustness** beyond the initial scope. These scenarios are designed to push the systems into operational conditions that are significantly more challenging than those adopted in the first phase.

There is a stronger emphasis on **comparing current performance metrics to the baseline** benchmarks established during the first campaign. This comparative analysis quantifies improvements and helps to identify persistent issues that require further refinement.

Based on insights from the first campaign, **the evaluation protocols are refined**. These adjustments result in a more targeted and effective validation process that directly addresses the limitations observed earlier.

To **enhance Human–AI Collaboration**, domain experts and operational staff are engaged continuously into the validation process, with refined operator training protocols and more frequent feedback

sessions. This ensures that human feedback and domain expertise continuously shape the testing environment, leading to iterative improvements in the AI solutions.

2.3.2.2 GOALS AND SCOPE

The primary objectives of the second validation campaign are following:

- Expose the AI systems to high-load scenarios and unexpected disruptions to evaluate their **scalability**, and overall **performance** in more complex operational environments.
- **Perform comparative evaluation** against the baseline benchmarks set during the first campaign. This process will quantify improvements in areas such as response time, accuracy, and operational impact while identifying any areas where challenges persist.
- **Enhance the collaboration between human operators and the AI systems** through refined training protocols and improved feedback mechanisms. This ensures that AI outputs are effectively integrated into operational decision-making and that operator expertise continuously informs system improvements.
- Verify that the AI systems maintain adherence to all **applicable regulatory standards** and **safety protocols**, even under increased complexity and live operational conditions.

2.3.2.3 EXPECTED OUTCOMES

At the conclusion of the second validation campaign, two deliverables will be produced.

D4.2 (AI4REALNET solutions' technical performance) will report the results of the AI solutions, considering mainly technical aspects (see §3-Technical performance and scalability and §4-Safety and robustness).

D4.3 (Evaluation of decision quality, trustworthy, and economics) will evaluate the proposed AI solutions considering:

- human user experience, acceptability and trustworthiness, and identification of required organizational changes (see §5-Social-technical decision quality),
- economics and regulation of the AI solutions (see §6-Economic and regulatory assessment).

2.4 ASSESSMENT

The validation campaigns will be assessed holistically, combining the quantitative and qualitative results from the execution of the evaluation protocol and the legal and economic evaluation. This has the advantage, that the results from the various aspects like the technical performance of AI models, the interaction with AI assistants, the robustness of suggestions made by AI, the potential impact of such Human-AI-interaction, and more can be discussed in relation to and depending on each other.

3. TECHNICAL PERFORMANCE AND SCALABILITY

This section details contribution of Task 4.1 to the evaluation, using a common template to define evaluation protocols, and a common template to define scenarios.

3.1 CONTEXT

Evaluating the technical performance of an AI assistant used in critical operation contexts is crucial: any error can lead to severe consequences, including system failures leading to loss of essential services for the community, or even endangering human lives (see risk assessment in §4 - Safety and robustness). Criticality of operations also requires that AI assistants can scale to the infrastructure's size properly.

Technical performance and scalability evaluations help maintain high standards of performance, build trust with human operators, and ensure the AI system contributes positively to the decision-making process as described in the Conceptual Framework of AI4REALNET project reported in deliverable D1.1.

3.2 EVALUATION PROTOCOLS

3.2.1 PERFORMANCE METRICS

Performance evaluation is focused on functional evaluation with a business and operation-oriented point of view.

AI systems support real-time network operations by integrating information and forecasted conditions, enabling corrective and preventive actions at various automation levels. Manual actions are emphasized in the power grid domain, while higher automation levels are considered for railway and ATM domains. Each domain's network structure helps inform solution strategies and constraints. Performance evaluation is therefore defined specifically per domain and divided into **Solution quality** and **Effectiveness**.

Solution quality refers to the overall excellence and suitability of the solutions provided by the AI assistant. It involves several factors, including the accuracy of the AI's responses and recommendations, ensuring they are correct and reliable.

Effectiveness refers to the AI assistant's ability to achieve desired outcomes and meet its intended goals. This includes the efficiency with which the AI performs tasks, completing them quickly and with minimal resource consumption, and the learning capacity, where the AI learns from interactions and improves over time, becoming more useful with continued use.

Evaluation protocols producing performance metrics are listed hereafter per domain.

3.2.1.1 EFFECTIVENESS

Air Traffic Management

Reduction in delay KPI-RF-027 (as described in annex 1) evaluates the percentage reduction in flight delays due to AI implementation. It highlights the AI's capability to improve operational efficiency and reduce delays.

System efficiency KPI-SS-032 (as described in annex 1) measures the efficiency of the system in delivering trustworthy solutions requiring less effort and time to deliver an appropriate response by the operator.

Power Grid

Assistant alert accuracy KPI-AF-008 (as described in annex 1) measures how accurately the AI assistant can forecast issues on the grid ahead of time. It directly impacts the AI's ability to achieve desired outcomes by preventing problems before they occur.

Network utilization KPI-NF-024 (as described in annex 1) assesses how well the grid network and its components are utilized. It indicates the AI's ability to optimize resource usage and improve overall network performance.

Total decision time KPI-TS-035 (as described in annex 1) measures the overall time needed to make decisions, including the time taken by both the AI assistant and human operator. Reducing decision time enhances operational efficiency.

Railway

Delay reduction efficiency KPI-DF-016 (as described in annex 1) quantifies how well the AI system reduces delays. It is a direct measure of the AI's ability to improve operational efficiency and achieve the goal of minimizing delays.

Punctuality KPI-PF-026 (as described in annex 1) measures the percentage of trains arriving on time or the aggregated delay in a scenario. It directly relates to the AI's ability to ensure timely operations.

AI Response time KPI-AF-029 (as described in annex 1) assesses the speed at which the AI system responds to disruptions or changes.

3.2.1.2 SOLUTION QUALITY

Power Grid

Carbon intensity KPI-CF-012 (as described in annex 1) evaluates the overall carbon intensity of the AI's action recommendations. It reflects the quality of the solutions provided by the AI in terms of their environmental impact, ensuring they are sustainable and responsible.

Topological action complexity KPI-TF-034 (as described in annex 1) gives insights into the complexity of topological actions utilized by the AI. High-quality solutions should avoid overly complex or hard-to-recover topologies, ensuring the solutions are practical and manageable.

Operation score KPI-OF-036 (as described in annex 1) includes the cost of blackouts, energy losses, and remedial actions. It reflects the overall quality of the AI's operational solutions, ensuring they are cost-effective.

Assistant adaptation to user preferences KPI-AS-068 (as described in annex 1) measures how well the AI assistant learns from the operator's choices. High-quality solutions are those that adapt to user preferences, providing personalized and relevant recommendations.

Railway

Network Impact Propagation KPI-NF-045 (as described in annex 1) assesses how small disruptions propagate under varying network scales.

3.2.1.3 USE CASES AND PROJECT GOALS

To ensure a proper functional evaluation, protocols must correspond to initial use cases' goals. Table 2 details which evaluation protocol allow to assess how initial use cases' goals are fulfilled.

Use case	Domain	Objective	Evaluation protocols
UC1.POWER GRID	Power Grid	Provide remedial action recommendations to allow the operator to safely managing overloads on the electrical lines.	KPI-AF-008 (Assistant alert accuracy) KPI-NF-024 (Network utilization) KPI-TF-034 (Topological action complexity) KPI-TS-035 (Total decision time) KPI-OF-036 (Operation score) KPI-AS-068 (Assistant adaptation to user preferences)
UC1.POWER GRID	Power Grid	Making the most of the renewable energies installed by exploring more in depth network topology optimization, instead of curtailement.	KPI-CF-012 (Carbon intensity) KPI-OF-036 (Operation score)
UC2.POWER GRID	Power Grid	Use an AI assistant in the real world while improving human trust.	KPI-AS-068 (Assistant adaptation to user preferences)
UC2.POWER GRID	Power Grid	Use an AI assistant in the real world while allowing for iterative human-AI refinements with human feedback and insights.	KPI-AF-008 (Assistant alert accuracy)

Use case	Domain	Objective	Evaluation protocols
UC1.Railway	Railway	Fully automate re-scheduling in railway operations to fulfill all offered services for the passenger.	KPI-PF-026 (Punctuality) KPI-AF-029 (AI Response time)
UC1.Railway	Railway	Fully automate re-scheduling in railway operations to minimize delays for the passenger.	KPI-DF-016 (Delay reduction efficiency) KPI-PF-026 (Punctuality) KPI-NF-045 (Network Impact Propagation)
UC2.Railway	Railway	Use AI-based methods to assist the human dispatcher in railway operations in re-scheduling train runs to fulfill all offered services for the passenger.	KPI-PF-026 (Punctuality)
UC2.Railway	Railway	Use AI-based methods to assist the human dispatcher in railway operations in re-scheduling train runs to minimize delays for the passenger.	KPI-DF-016 (Delay reduction efficiency) KPI-PF-026 (Punctuality) KPI-NF-045 (Network Impact Propagation)
UC1.ATM	ATM	Partially or fully automate the sectorisation process to assist or replace the ATC supervisor in deciding when and how to split and merge sectors to balance the workload of tactical ATCOs.	KPI-SS-032 (System efficiency)
UC2.ATM	ATM	Provide advice to ATCO about deviations to avoid the activated military area.	KPI-RF-027 (Reduction in delay)

Use case	Domain	Objective	Evaluation protocols
UC2.ATM	ATM	Provide advice to ATCO about deviations to review of the sectorisation plan.	KPI-SS-032 (System efficiency)

TABLE 2 – RELATION OF TECHNICAL PROTOCOLS TO USE CASES' OBJECTIVES

On a project level, evaluation of Long Term Impacts[3] (LTEI) will be measured in D4.2 where results of the evaluation will be reported. Table 3 thus details how protocols allow to perform this evaluation.

LTEI	Domain	Description	Evaluation protocols
(LTEI1)KPI$_{S-1}$	Power Grid	15%-20% reduction in renewable energy curtailment due to optimal exploration of network flexibility with AI	KPI-OF-036 (Operation score)
(LTEI1)KPI$_{S-2}$	Power Grid	20%-30% avoided electricity demand shedding	KPI-OF-036 (Operation score)
(LTEI1)KPI$_{S-3}$	Railway	10% increase in punctuality in long-range traffic 5% increase in punctuality in regional traffic (with realistic disturbances)	KPI-PF-026 (Punctuality)
(LTEI1)KPI$_{S-4}$	ATM	3-6% improvement in flight capacity and mile extension	KPI-RF-027 (Reduction in Delay)

TABLE 3 – RELATION OF TECHNICAL PROTOCOLS TO THE PROJECT'S LTEIS

3.2.2 SCALABILITY METRICS

Measuring the scalability of an AI assistant involves evaluating its ability to handle increasingly complex scenarios without compromising performance. Two main aspects of scalability are considered:

- Scalability at training time, which measures how much more time and resources are needed to train a model on larger scenarios,
- Scalability at testing time, which evaluates the ability of the agent to handle larger scenarios while evaluating potential additional time costs for the decision making.

Training time scalability KPI-AF-050 (as described in annex 1) considers as a key metric the time to reach a performance threshold. This metric allows comparing the time needed for different methods

[3] See Description of the Acction

to reach the same level of performance and can be expressed in seconds (wall clock time[4]), or the number of discrete time steps or episodes in the environment. This time-to-threshold is then measured in two different sets of circumstances, thus resulting in two curves: first, we consider scenarios of growing size, and measure how the time to threshold varies as a result (on standardized hardware, to keep the results comparable). Secondly, we consider different hardware setups and again measure time-to-threshold as dependent variable (on a standardized scenario). This last analysis shows how much the method under consideration can benefit from increased resources (with time-to-threshold measured as wall clock time in seconds).

Test time scalability KPI-AF-051 (as described in annex 1) considers also scenarios of increasing complexity, but with models that are already trained to an acceptable performance, and measures how long it takes these models to decide on an action or suggestion (measured in seconds, on a standardized hardware set-up). Since trade-offs between scalability and performance level can be expected (higher-performing model might be slower), the quality of decisions is measured to allow easy comparison between methods. This KPI thus also results in two curves: time requirement and decision quality as a function of instance complexity.

These scalability metrics are defined in the same way for all domains, taking advantage of the shared notion of "networked infrastructure system", which can always be associated with a sizing metric (e.g. number of nodes and links).

3.3 OPERATIONAL TESTING SCENARIOS

Starting from the use cases defined in the Deliverable D1.1 – Conceptual Framework, realistic and representative operational testing scenarios have been defined for each domain and will be used throughout the validation campaigns.

These scenarios have been defined to ensure that all the evaluation dimensions of Deliverable D4.1 – Evaluation and test protocols are covered (see Table 4 – relation of scenarios to): Technical performance and scalability, Robustness and safety of the AI solutions, Optimal balance between AI and human.

Evaluation dimension	Scenarios
Technical performance and scalability	**ATM**: Military Reservation OPSCE-UC2.A-1-011 **Power Grid**: Remedial actions OPSCE-UC1.P-1-001 **Railway**: Reactive Re-Scheduling OPSCE-UC2.R-1-007, Proactive re-scheduling OPSCE-UC2.R-3-009

[4] Elapsed real time, real time, wall-clock time, wall time, or walltime is the actual time taken from the start of a computer program to the end: difference between the time at which a task finishes and the time at which the task started.

Evaluation dimension	Scenarios
Robustness and safety of the AI solutions	**ATM**: Adverse weather conditions OPSCE-UC1.A-1-012 **Power Grid**: Real world conditions OPSCE-UC2.P-1-003 **Railway**: Re-Scheduling at the occurrence of infrastructure malfunction OPSCE-UC1.R-1-004, Emergency response to weather adverse weather conditions OPSCEUC1.R-2-005, Closure of a large station OPSCE-UC1.R-3-006
Optimal balance between AI and human	**ATM**: Military Reservation OPSCE-UC2.A-1-011, Weather Perturbations OPSCE-UC1.A-1-012 **Power Grid**: AI assistant learns OPSCE-UC1.P-2-002 **Railway**: Co-learning for reactive re-scheduling OPSCE-UC2.R-2-008, Co-learning for proactive re-scheduling OPSCE-UC2.R-4-010

TABLE 4 – RELATION OF SCENARIOS TO EVALUATION DIMENSIONS

Each scenario is instantiated using one or more datasets, that are either built within the range of ordinary conditions (i.e. within the distribution of input data space) or are specifically designed to test the limits of the systems (thus outside or at the boundaries of input data space).

The possibility to instantiate digital environments by using real data to simulate realistic conditions is different per domain. While ATM and Railway domains can easily reuse real open-sourced data, using real data from Power Grid is more difficult, especially due to confidentiality issues for generation units: depending on legal constraints, it is therefore envisaged that real data could be used as confidential data that won't be open sourced.

Air Traffic Management

Starting from UC1.ATM (Airspace sectorisation), **Adverse weather conditions** scenario **OPSCE-UC1.A-1-012** (as described in annex 2) focuses on a new sectorization plan in response to adverse weather conditions (e.g. Volcanic ashes). The AI-based system (**Airspace sectorization assistant**) detects the perturbation, recalculates new sectorization plans and applies an optimized solution to avoid the affected airspace.

Starting from UC2.ATM (Flow & Airspace management), **Military Reservation** scenario **OPSCE-UC2.A-1-011** (as described in annex 2) focuses on a new sectorization plan and/or routing in response to a sudden airspace reservation. The AI-based system (**Flow & Airspace management assistant**) detects the disruption, recalculates flight routes and the need for a new sectorization plan (if possible) and new traffic load forecasts and applies an optimized solution to minimize delays while complying with operational requirements (e.g. safety margins).

Power Grid

Starting from UC1.Power Grid (AI assistant), **Remedial actions** scenario **OPSCE-UC1.P-1-001** (as described in annex 2) describes how the AI assistant provides the human operator with warnings in anticipation to possible disruptions along with recommendations for remedial actions .

Starting from UC1.Power Grid (AI assistant), **AI assistant learns** scenario **OPSCE-UC1.P-2-002** (as described in annex 2) describes how the AI assistant updates its list of recommendations with actions that were performed by the human operator under different prevailing operating conditions.

Starting from UC2.Power Grid (Sim2Real), **Real world conditions** scenario **OPSCE-UC2.P-1-003** (as described in annex 2) focuses on AI assistant's robustness to bad or low-quality data.

Railway

UC1.Railway refers to human-AI interactions to augment decision-making and UC2.Railway refers to integrated autonomous AI-driven decision systems: with UC1.Railway, AI assistant solves all problems independently. However, the human can cancel the execution, and plays an "active supervision" role.

Starting from UC1.Railway (Automated rescheduling), **Re-Scheduling at the occurrence of infrastructure malfunction** scenario **OPSCE-UC1.R-1-004** (as described in annex 2) focuses on the automated railway management system facing a challenge when a sudden infrastructure malfunction occurs (trigger event). This requires the selection of the most appropriate response strategy and its immediate implementation to ensure continued service delivery and minimize performance loss.

Starting from UC1.Railway (Automated rescheduling), **Emergency response to adverse weather conditions** scenario **OPSCE-UC1.R-2-005** (as described in annex 2) deals with sudden weather challenges, such as extreme weather conditions, impacting railway operations.

Starting from UC1.Railway (Automated rescheduling), **Partial closure of a large station** scenario **OPSCEUC1.R-3-006** (as described in annex 2) addresses the challenge of rescheduling in case of the partial closure of a major station.

Starting from UC2.Railway (AI-assisted human rescheduling), **Reactive Re-Scheduling** scenario **OPSCE-UC2.R-1-007** (as described in annex 2) shows the reactive re-scheduling by the human-AI team once an unexpected disruption has already occurred.

Starting from UC2.Railway (AI-assisted human rescheduling), **Co-learning for reactive re-scheduling** scenario **OPSCE-UC2.R-2-008** (as described in annex 2) shows the co-learning process initialized by the reactive re-scheduling by the human-AI team once a deviation or disturbance has already occurred.

Starting from UC2.Railway (AI-assisted human rescheduling), **Proactive re-scheduling** scenario **OPSCE-UC2.R-3-009** (as described in annex 2) shows the proactive re-scheduling by the human-AI team upon detection of weak signals before a disruption occurs.

Starting from UC2.Railway (AI-assisted human rescheduling), **Co-learning for proactive re-scheduling** scenario **OPSCE-UC2.R-4-010** (as described in annex 2) shows the co-learning process initialized by the proactive re-scheduling by the human-AI team.

4. SAFETY AND ROBUSTNESS

This section details the contribution of Task 4.2 to the evaluation, using a common template to perform a risk assessment and to define evaluation protocols.

4.1 EUROPEAN CONTEXT FOR AI SAFETY

4.1.1 AI ACT: ROBUSTNESS, RELIABILITY, AND RESILIENCE

The European Union's AI Act emphasises the need for accuracy, robustness, and cybersecurity throughout their lifecycle. Article 15 ("Accuracy, robustness and cybersecurity") demands *"measures to prevent, detect, respond to, resolve and control for attacks trying to manipulate the training data set (data poisoning), or pre-trained components used in training (model poisoning), inputs designed to cause the AI model to make a mistake (adversarial examples or model evasion), confidentiality attacks or model flaws"*. Moreover, in point (27) of the AI Act, it is also mentioned *"Technical robustness and safety means that AI systems are developed and used in a way that allows robustness in the case of problems and resilience against attempts to alter the use or performance of the AI system so as to allow unlawful use by third parties, and minimise unintended harm"*, which, in addition to technical robustness (a key requirement in high-risk systems), it also introduces the notation of resilience. This is reinforced in Article 15 with *"High-risk AI systems shall be as resilient as possible regarding errors, faults or inconsistencies that may occur within the system or the environment in which the system operates, in particular due to their interaction with natural persons or other systems"*.

At this point, it is crucial to introduce the definitions of robustness and resilience from the conceptual framework in Deliverable D1.1 (Bessa, et al., 2024), as these concepts are central to the discussions throughout the paper. **Technical robustness** is a system's ability to maintain its performance level under natural or adversarial perturbations. It can be local (specified concerning a sample input) or global (guarantees that hold deterministically over all possible inputs) (ISO/IEC 24029-2:2023). Considering the complexity of the systems at hand in AI4REALNET, local robustness is easier to specify and verify. **Resilience** is the ability of an AI system to prepare for and adapt to changing conditions and withstand and recover (i.e., return to a "normal" state) rapidly from natural or adversarial perturbations or unexpected changes (EU-U.S. Terminology and Taxonomy for AI). Here, it is important to highlight the notion of recovery in resilience.

Although not mentioned in the AI Act, reliability is also defined in the project conceptual framework as focused on consistent performance aligned with the underlying data distribution in standard operating environments. This definition is strongly related to out-of-domain data, and it means that the AI system should perform similarly on any test sets/periods if they are from the same distribution.

The technical aspects of measuring the appropriate levels of robustness and other relevant metrics are expected to be developed in cooperation with relevant stakeholders such as standardization bodies, benchmarking authorities, and research and innovation projects like AI4REALNET.

4.1.2 CONCEPTS FROM STANDARDIZATION WORK GROUPS

While the AI Act discussed in previous subsection sets these critical requirements, it lacks a specific methodology for quantifying robustness and does not give high relevance to the concept of resilience. Therefore, standardisation organisations like the International Organization for Standardization (ISO), the International Electrotechnical Commission (IEC), and the Institute of Electrical and Electronics Engineers were mandated to create a family of standards that cover this gap.

One notable example is the ISO/IEC 24029-2 standard (*Assessment of the robustness of neural networks. Part 2: Methodology for the use of formal methods*) which covers different formal methods (ISO/IEC TR 24029-1:2021) to assess the robustness of artificial neural networks (ANN) and defines properties such as stability, sensitivity, relevance, and reachability. However, as analysed by the Joint Research Center (Garrido, et al., 2023), this standard is mostly applicable to classical AI applications (e.g., classification, regression), although in the literature there are already approaches for reinforcement learning ((Fulton, et al., 2018); (Corsi, et al., 2021)). Also, it does not propose performance metrics to assess these properties. Nevertheless, the properties in this standard were used to derive evaluation metrics in Section 4.3.3 and the terminology/taxonomy to describe the methodology in Section 4.3.1. It is also important to mention that, as highlighted in ISO/IEC 24029-2, formal methods do not replace other means of verification and validation, as the creation of adversarial instances to train and validate (pre-trained models in testing-time) AI-based agents – as foreseen in AI4REALNET.

The standard ISO/IEC 23894:2023 (*Guidance on risk management*) provides guidance for managing risks associated with AI and mainly applies to organisations involved in developing, deploying, or using AI-based products, systems, and services; it integrates the principles of risk management from ISO 31000:2018 into AI-related activities. It emphasises stakeholder engagement and interdisciplinary expertise to address the unique risks introduced by AI, such as bias, explainability challenges, and dynamic learning behaviors. It also outlines a framework for integrating risk management across the AI lifecycle, from inception and design to operation and retirement. Specific guidance is provided for risk identification, assessment, treatment, and monitoring, emphasizing traceability and continual improvement. Furthermore, the Massachusetts Institute of Technology published in 2024 the AI Risk Repository (Slattery, et al., 2024), which captures more than 700 risks extracted from 43 existing frameworks and classifications of AI risks, as well as a causal taxonomy of AI risks (e.g., classifies how, when, and why these risks occur) and a domain taxonomy of AI risks (e.g., classifies these risks into seven domains and 23 subdomains. However, the risks in the database are not domain-specific and lack the detailed contextualization and specificity required to address the challenges and vulnerabilities of critical infrastructures. Therefore, the risk identification and assessment process applied in Section 4.2 and detailed in Annex 3 was based on the definitions and processes included in the ISO 31000:2018 and ISO/IEC 23894:2023 standards to align the risk assessment methodology in AI4REALNET with the current European standards.

The formalization of risks also follows the multi-component framework proposed in the AI4REALNET conceptual framework (Bessa, et al., 2024), which is based on consolidated methodologies in disaster

risk reduction – e.g., the Intergovernmental Panel on Climate Change risk framework (Reisinger, et al., 2020) – and allows to cover the caveats of EU AI Act, which lack representation of the risk sources (hazards) and factors, that make involved actors vulnerable to these hazards. Including these points in the assessment allows a more precise understanding of the effectiveness of mitigation strategies later. In the later stage of robustness and safety assessment, separating risk into risk components allows a more accurate comparison of different risk sources and quantification of the severity of their outcomes for the system and stakeholders.

4.2 RISK IDENTIFICATION AND ASSESSMENT

This subsection summarizes the risk assessment conducted at the beginning of WP4, which followed the methodology and terminology described in Annex 1. This risk assessment covers different types of risks, but the focus in this section was aligned with Article 15 ("Accuracy, robustness, cybersecurity") of the AI Act, meaning technical risks that can influence the AI system and are associated with errors, faults, or inconsistencies that may occur within the system or the environment in which the system operates.

An initial risk assessment was conducted during Month M15 of the project, with the detailed results provided in Annex 3. A summary related to Article 15 is presented below.

Air traffic management

Attacks on the AI model are rated as low due to robust cybersecurity measures in the ATM infrastructure. Successful attacks would likely cause minor disruptions, such as incorrect pre-tactical routing or vectorisation decisions, but operational systems would remain unaffected due to mitigation protocols and rapid restoration mechanisms. The exposure level is medium as the model impacts several ATM components.

For reward/loss functions, risks are also low due to their internal nature and strong security protections. However, when this occurs, it could learn inappropriate policies, leading to suboptimal or unsafe operational decisions, and resource misallocations could occur. The range of impact could span from minor operational disturbances to severe safety incidents.

The action/output space is at moderate risk since the outputs of the AI system will be submitted to human operators, and they may be more accessible to attackers compared to internal components like the model or reward function. If the communication channels are not fully secured, the risk of such perturbations increases. Perturbations here could disrupt Tactical Air Traffic Controller (ATCO) scheduling and sectorisation, with potential safety implications. While restoration is typically fast, the exposure level is high, as outputs directly influence downstream systems, operational staff, and decision-making processes.

The state/input space is the most vulnerable, with a risk rating of moderate to high due to heavy reliance on external data sources such as surveillance systems, databases, and user inputs. Attacks can manipulate or corrupt inputs, leading to erroneous decisions, scheduling conflicts, and operational disruptions. Vulnerabilities arise from unsecured communication channels, lack of validation

mechanisms, and dependency on external data. Exposure is high, as these inputs affect a wide range of systems, including safety mechanisms, control systems, and human operators.

Power grid

Low-risk ratings (i.e., frequency) are assigned to robust cybersecurity measures protecting internal components like the AI model and reward functions. However, adversarial attacks at the state/input space are rated as a medium since the AI system relies on a multitude of input data sources on the electric system—including sensors, external databases, communication networks, and user inputs, same as, e.g. supervisory control and data acquisition (SCADA) systems—there is a tangible risk that these inputs can be intentionally manipulated or inadvertently corrupted.

The potential hazards include manipulated model outputs, compromised system security, and degraded model performance, which could lead to system imbalances or line overload (potentially blackouts). The AI system shall be safeguarded within the same cybersecurity and restoration protocols as the other IT assets to ensure no discrepancies in security in the operations' tooling environment. While measures are in place to mitigate these vulnerabilities, all AI models remain exposed to these risks, though the extent of impact remains difficult to quantify.

Another identified risk is changes in the underlying data distribution over time, rated as medium. The context of the energy transition is bringing major changes into the whole EU electric system, which affects the type of generation/load units connected and stakeholders' behaviours. In addition, these units tend to be more climate-dependent (RES, but also hydro, nuclear), which in turn can shift data patterns over time more rapidly than before due to climate change. The latter can also bring important changes that are more limited in time, such as extreme events. This can affect the AI model in terms of decreased accuracy, unreliable predictions, and the need for frequent retraining. The AI model needs frequent retraining on live data, and the development of the AI model should decrease its sensibility to data distribution shifts.

Railway network

Attacks on the AI model are rated as low risk due to advanced cybersecurity measures in place. However, successful attacks could cause operational disruptions, safety risks, financial losses, and reputational damage. The AI model's secure environment and rapid recovery mechanisms minimise its vulnerability. For reward/loss functions, the risk remains low, as these components are internal and well-protected. Nevertheless, insider or external threats could exploit these functions, leading to incorrect decision-making, safety hazards, or resource misallocations. While the AI system's vulnerability here is low, the centrality of these functions means their compromise could significantly affect the entire system.

The action/output space has a moderate risk rating, as it is more accessible than internal components. Attacks here could disrupt operations, cause scheduling conflicts, and jeopardise safety. Restoration is generally quick with effective monitoring, but the potential for significant disruption requires robust output validation and secure communication protocols.

The state/input space poses the highest risk, with a rating of moderate to high. The system's dependence on numerous external data sources—such as sensors, communication networks, and user inputs—creates opportunities for manipulation. Attacks on inputs can severely disrupt operations, misroute trains, and compromise safety. Vulnerability increases with unencrypted data transmission and inadequate input validation. Addressing these risks involves securing input channels, rapidly filtering corrupted data, and mitigating cascading effects.

Overall analysis

A common aspect of these three infrastructures is that the critical technical risk source is perturbations in the state/input space. These perturbations are more frequent (e.g., information collected from different internal and external data sources) and have a higher potential impact, as they cannot be mitigated via model replacement or retraining.

Another risk mentioned in the power grid – and that leads to one specific use case, see (Bessa et al., 2024) for more details – but that could be applied to other infrastructures is out-of-domain data, particularly when digital environments are used for the initial training of the AI-based decision system, and significant differences might occur between digital and real environment.

Finally, from the application of the Assessment List for Trustworthy Artificial Intelligence (ALTAI) tool to the project's use cases (Bessa, et al., 2024), the following conclusions were also derived for Requirement #2 Technical Robustness and Safety[5]:

- In the power grid domain, data quality issues such as noise, missing data, and adversarial attacks on the state vector—the characterisation of the operating context—can lead to incorrect decisions. Environmental threats, particularly extreme weather conditions, further complicate AI-driven decision-making. High epistemic uncertainty, resulting from a lack of representative training data, increases the risk of erroneous forecasts. Furthermore, the misuse of AI systems can lead to incorrect decisions by human operators. It is crucial to implement automatic mechanisms capable of detecting data and AI model shifts to address these risks. Stress testing should be performed to assess the resilience of AI models, considering perturbations in the state vector, model parameters such as weights, and system outputs.
- For railway networks, while no specific risks were initially identified, the critical nature of the infrastructure managed by AI systems underscores the necessity of fault tolerance. Given the potential consequences of AI failure, ensuring that AI-driven decision-making remains reliable under unexpected conditions is essential. Additionally, performance monitoring is particularly important when AI systems rely on online learning agents, whose behaviour may change over time.
- In air traffic management, AI systems are prone to cyber intrusions and data corruption, making stability and reliability fundamental requirements. These AI systems must function correctly under both normal and unexpected conditions. Reinforcement learning (RL)

[5] https://ai4realnet.eu/wp-content/uploads/2024/08/D1.1-ALTAI_Summary.pdf

algorithms, commonly used in this domain, must adapt to data shifts, traffic load variations, and evolving operational conditions. However, continuous learning should not introduce unintended risks. To ensure robust AI performance, algorithmic changes must undergo rigorous testing before being deployed in production environments. Furthermore, users must be aware of any software state changes to prevent disruptions in ATM operations.

4.3 EVALUATION PROTOCOLS

4.3.1 METHODOLOGY

The envisioned methodology in the Grant Agreement fundamentally consists of how to create adversarial datasets – see Section 4.3.2 – that encompass the critical risks identified in Section 4.2. Evaluation metrics from Section 4.3.3 are used to test the robustness, reliability, and resilience of the AI-based functionalities, considering the data-driven nature of AI and the fact that security and data breaches of AI systems can progressively influence future decision quality due to continuous training and updating of the underlying AI system. Besides, these adversarial examples will also evaluate how the developed solutions deal with real-world conditions, such as missing, wrong, or delayed measurements, data distribution shift, etc., to understand how the step from simulated environments to real-world environments impacts the performance of the solutions. These perturbations can be applied during the training of the AI-based system to improve its performance, named 'training-time', as well as during the system's operation for conformity assessment, named 'testing-time'.

The creation of adversarial datasets will be integrated into/with the digital environments since the domain of robustness assessment is composed of the attributes in the state vector, reward, and output vector that is available in the digital environments Grid2Op, Flatland, and BlueSky, as well as the implemented use cases/tasks. As mentioned by ISO/IEC 24029-2, the perturbations can be beneficial or negative, meaning that perturbations do not necessarily lead to the worst performance of the AI agent during testing-time. Moreover, as defined in ISO/IEC 24029-2, these perturbations could be intentional (e.g., adversarial attack) or unintentional (e.g., missing or erroneous measurement, change in the environment), and both options can be considered to evaluate a system.

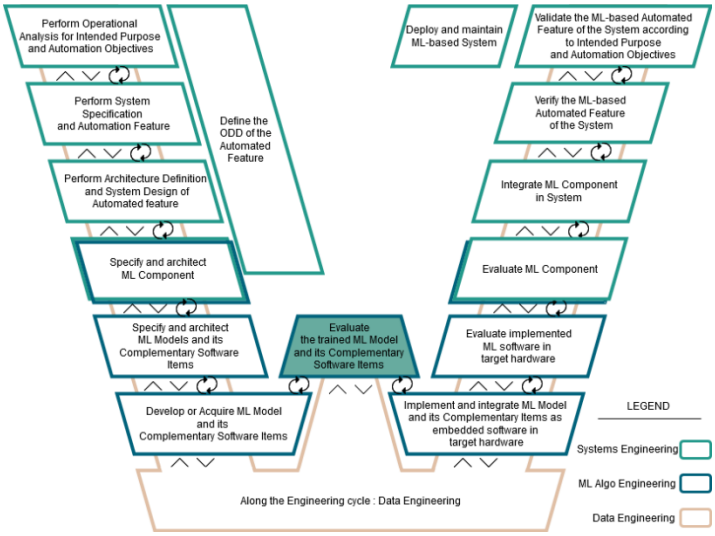


FIGURE 8 – CONFIANCE.AI END-TO-END APPROACH FOR THE ENGINEERING OF CRITICAL TRUSTWORTHY ML-BASED SYSTEMS

As an example of a methodology for evaluating the trustworthiness of AI-based systems in critical environments, the Confiance.ai research program[6] revisited classical engineering disciplines, including Systems Engineering and Software Engineering, to effectively integrate ML technology into these systems. It has thoroughly formalized the engineering processes required for ML-based critical systems, which can be accessed through an interactive interface here[7]. It is based on the drafts of standards that address the subject of machine learning systems (e.g., ISO/IEC DIS 5338, AS 6983) to provide the overall structure of an end-to-end engineering approach. Figure 8 shows the main engineering steps of this end-to-end method (Awadi, et al., 2024), organised according to the traditional "V" cycle ("W" cycle at the ML model level). The program also emphasises that trustworthiness must be ensured at every stage of developing ML-based systems (e.g., model development, training and evaluation).

It is emphasized that the trustworthiness attributes can be assessed only if the Operational Design Domain (ODD) is clearly defined (first step in the scheme above). The ODD refers to operating conditions under which a given AI-based system is specifically designed to function as intended, in line with its intended purpose. In other words, it defines conditions in the environment where the AI subs-subsystem shall operate. Hence, an understanding of the real environment and its matching to the training distribution is required to effectively predict the performance or the risks in an operational context.

Once an ML model has been developed (built, configured, and trained), it should be evaluated to guarantee that the required trustworthiness properties, such as robustness to adversarial perturbations, are satisfied. The IEEE Glossary of Software Engineering (ANSI/IEEE Std 729-1983) defines robustness as the degree to which a system or component can function correctly in the presence of invalid inputs or stressful environmental *conditions*. Moreover, in (ISO/IEC TR 24029-

[6] https://www.confiance.ai/

[7] https://bok.confiance.ai

1:2021), it is *the ability of an AI system to maintain its level of performance under any circumstances*. The robustness of a system relates to the question of whether the system can be trusted to perform well its intended purpose in the envisioned ODD.

There are mainly two techniques that could be considered to evaluate the robustness of ML-based models: a) *Empirical robustness evaluation* by assessing the robustness of a trained model against a set of perturbations, where the deviation of the model performance from its nominal behaviour is measured. This provides initial insights but may lack reliability as it only covers a subset of possible scenarios; b) *Formal evaluation of robustness* consisting of verifying mathematically the model's robustness, expressed using formal properties, by accessing the internal structure of the model and, in some cases, the dynamic (state-space) model of the environment. Given the cost of formal verification, Confiance.ai program recommends starting with empirical robustness evaluation and performing robustness formal verification only once the empirical evaluation obtains satisfactory results (see Figure 9). As can be seen in this scheme, one of the requirements enabling the assessment of robustness is the design and creation of adversarial datasets.

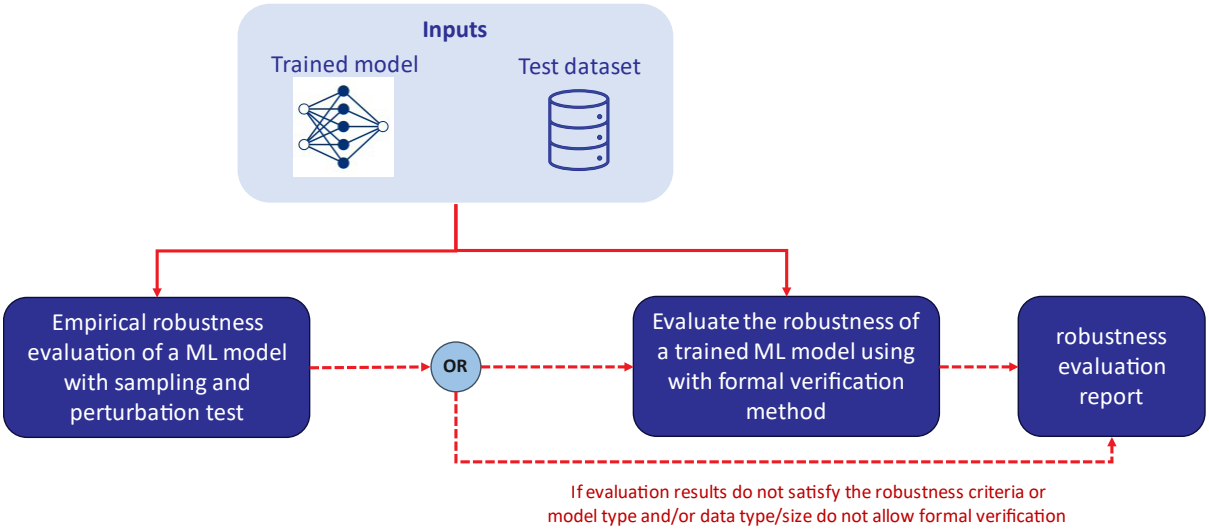


FIGURE 9 – ROBUSTNESS EVALUATION PROTOCOL OF ML MODELS

To ensure the overall applicability of the robustness evaluation protocol, certain prerequisites must be met:

- A trained ML model to be evaluated for robustness. The ML-based model will rely on a RL paradigm and address the control problem in different critical infrastructure application domains and use cases (e.g., power grid control, railway management, and air traffic management),
- The robustness requirements applicable to the selected ML model serve as the reference against which the robustness evaluation of the trained ML model should be conducted,
- The robustness evaluation strategy for the ML model,
- A test dataset should be designed for the unit evaluation of the selected ML model.

The robustness evaluation also requires sufficient specificity regarding the training objectives of the AI agent, particularly in the context of multi-objective RL. Changes in agent behaviour induced by changing the relative weights of rewards, equivalent to moving along an optimized pareto front, does not constitute a test of robustness, and can interfere with evaluating the effects of other perturbations. Therefore, in the empirical robustness evaluation of multi-objective agents, it is recommended that a fixed reward structure is assumed during robustness tests.

4.3.1.1 EMPIRICAL ROBUSTNESS EVALUATION

The *empirical robustness* of an ML-based approach or system could be evaluated using a step-by-step guide depicted in Figure 10. It starts by gathering the required information concerning the dataset, the model, the robustness properties required for a specific scenario and context, and finally, a robustness evaluation strategy to be adopted. The next step involves applying a specific type of perturbation to the provided dataset on which the model should operate. Next, the robustness evaluation is performed by comparing the model prediction using the perturbed dataset and considering its nominal behaviour using the unperturbed dataset. As the result of the evaluation, two scenarios could happen: If the difference is within the maximum permissible threshold, the robustness requirement is satisfied, indicating the model can handle the given perturbation level; If the difference exceeds the threshold, the requirement is not met, but the study may continue by reducing perturbation intensity to measure robustness within the acceptable tolerance. If the ML model fails to meet the robustness requirement, the evaluation will be analysed to understand the gap. Based on this analysis, decisions will be made to adjust the development strategy or renegotiate requirements to improve the robustness of ML-based models.

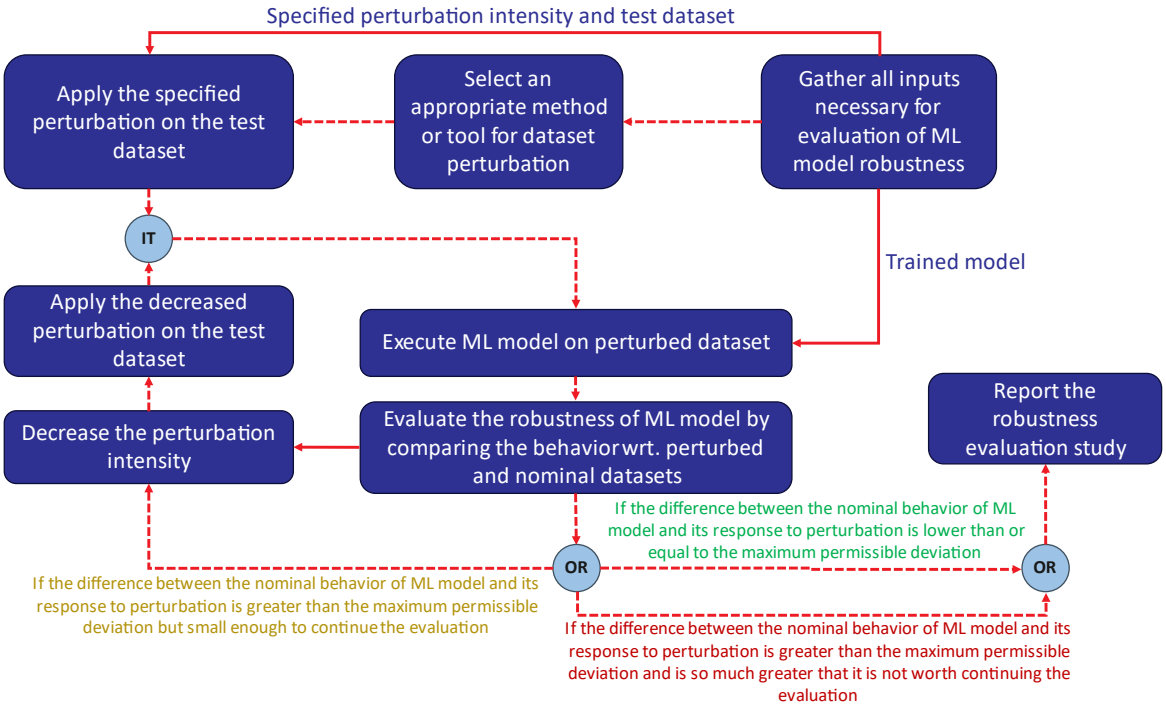


FIGURE 10 – EMPIRICAL ROBUSTNESS EVALUATION PROTOCOL AGAINST PERTURBATION OF THE TEST DATASET

To formalise the problem, given a model f for decision-making, an adversarial attack (input) $\acute{x}$, to be effective, should verify the following conditions for $\delta > 0$:

$$\exists \acute{x} : d(x, \acute{x}) \leq \delta \text{ and } f(x) \neq f(\acute{x}), \tag{4.1}$$

where d is a choice of distance between inputs. As such, the minimal distance δ^* such that (4.1) holds is exactly given as the distance between $f(x)$ and the decision boundary of f. Validating this property means crafting adversarial examples. This is usually formalised as a problem of finding

$$\underset{\acute{x} \in B(x)}{\operatorname{argmax}}\, g(\acute{x}) \tag{4.2}$$

where $g(\acute{x})$ measures the successfulness of the attack, $B(x)$ is a chosen neighbourhood of the reference input x indicating the feasible perturbation range. Creating such a counterexample does not directly give information on the overall behaviour of the model. Evaluating adversarial robustness empirically means quantifying this behaviour. The usual way of producing such empirical local characterisation is to have an algorithmic way of translating a clean input into an adversarial one so that one can produce an adversarial copy of a test set and then measure the model's performance drop when tested on such an adversarial test set.

4.3.1.2 FORMAL VERIFICATION OF ROBUSTNESS

Similarly, the formal verification of robustness method has its own procedure to assess the robustness of an ML-based approach using mathematical verification (see Figure 11). It starts by gathering the required inputs, i.e., expressing ML robustness requirements as formal properties, defining an evaluation strategy, pre-trained models for evaluation, a test dataset, and a state-space mathematical model of the environment in RL. Next, the formal properties should be verified by applying the formal method to the trained model for each element of the test dataset. Finally, the results of the formal analysis could be analysed to investigate the robustness of compliance requirements.

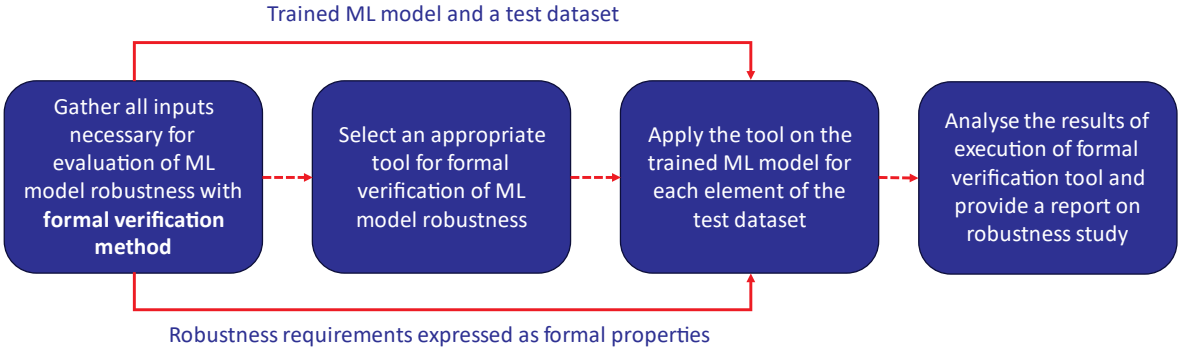


FIGURE 11 – FORMAL EVALUATION OF ROBUSTNESS PROTOCOL

Formally proving adversarial robustness does not suffer from the first problem (previous sections have rigorously mathematically defined this notion, which is ample formal specification). The scaling problem, however, remains a hurdle today despite a considerable amount of research on the matter. The current state-of-the-art still makes formal verification a worthwhile, stronger version of empirical adversarial robustness evaluation. In a nutshell, instead of selecting a set of inputs I, then generating

(an obviously limited number of) adversarial examples I' within a given neighbourhood (a distance ε) of these inputs and using I' set to evaluate the model, formal verification allows for direct evaluation of the entire neighbourhood of the inputs in I within ε. It still does not cover all possible inputs (only I), but it is strictly more exhaustive than the generation of a limited number of adversarial inputs. In the case where the AI agent is formulated in a multi-objective fashion, a complete evaluation of robustness would also consider the impact of perturbations on the baseline pareto front established by the agent in the nominal scenario.

In this project, we consider only the empirical evaluation of robustness by defining different strategies for the generation of adversarial datasets and a set of metrics (generic and domain-specific) through the developed models.

4.3.2 GENERATION OF (ADVERSARIAL) DATA PERTURBATIONS

This section outlines the specification of perturbation agents responsible for generating (adversarial) datasets used to evaluate robustness and resilience empirically. It considers generic, environment-agnostic perturbation agents based on gradient estimation, RL, and action perturbations designed to represent intentional manipulations—such as cyberattacks—that evade detection by human experts as well as choices by the human operator who might choose to accept an AI agent's suggestions or ignore them. Additionally, domain-specific perturbation agents are included to incorporate domain knowledge (e.g., identifying which variables to perturb) and to simulate natural or unintentional adversarial perturbations.

As discussed in the section 4.2, the adversarial perturbation agents generate adversarial examples in the input space of the AI system, as this is the most probable perturbation in AI systems installed in control rooms of critical infrastructures. While the agent cannot directly modify the actual state of the network (environment), it can alter input data measurements, thereby influencing the AI system's outputs. Furthermore, the perturbation agent can be employed during training-time and testing-time.

Generic perturbation agents

Four generic perturbation agents are introduced here; one leverages gradient-based sensitivity information from the ANN, termed the "gradient estimation perturbation agent," the second employs a RL algorithm, known as the "RL-based perturbation agent,". The third and fourth are similar in nature, termed "action perturbing agent" and "communication perturbing agent". The former provides randomized human-operator like perturbations to the suggested actions, while the latter introduces simulates external variance to inter-agent communication.

Gradient estimation perturbation agent. Creates an adversarial example to minimise the output of the optimal action in the policy of the AI agent. This process is repeated during every step of the digital environment. These adversarial examples are created using the combination of gradient estimation and projected gradient descent, also used in (Chen et al., 2021), which is why it is called the gradient estimation perturbation agent. The gradient estimation is needed since this work is focused on black-box attacks, and it is done using Eq. 3.1. In this Equation, L(s) is the target function that the attacker wants to change. The adversarial examples this agent creates are untargeted, meaning that it does

not matter which action the AI agent takes as long as it is not optimal. Therefore, L(s) is defined as the value corresponding to the optimal action in the policy of the AI agent and the goal is to minimise this value to make the action less attractive. The vector e_i in the Eq. has a 1 in the i-th position and 0 everywhere else.

In Eq. (4.3), *L(s)* is the target function for the adversarial example. Since the goal in this work is to make the AI agent perform a certain action, *L(s)* is defined as the value corresponding to the action in the policy of the AI agent. The vector e_i in Eq. (4.3) has a 1 in the i_{th} position and 0 everywhere else.

$$g_i = \frac{\partial L(s)}{\partial s_i} \approx \frac{L(s + 0.01 \cdot e_i) - L(s - 0.01 \cdot e_i)}{0.02} \tag{4.3}$$

These estimated gradients can be computed for each value of an observed state in the digital environment and can be used in the following projected gradient descent algorithm. The algorithm is summarized in Figure 12.

Algorithm 1: Gradient estimation perturbation agent

Data: s, W, ζ, ξ

$s^{adv} \leftarrow s$;
for $w = i, \ldots, W$ **do**
  Compute gradients $g(s^{adv})$;
  $s^{adv} \leftarrow s^{adv}(1 + \zeta sign(g))$;
  Clip s^{adv} to keep it between $s(1 - \xi)$ and $s(1 + \xi)$;
end

FIGURE 12 – ALGORITHM OF THE GRADIENT ESTIMATION PERTURBATION AGENT

In this algorithm, s is the vector representation of the state that needs to be perturbed, and s^{adv} is the adversarial example. Additionally, *W* is the number of iterations in the projected gradient descent, ζ is the step size in each iteration and ξ is the maximum perturbation.

RL-based perturbation agent. It actively aims to change the behaviour of the AI agent with as little perturbation as possible. This agent is based on the one developed by (Garcia et al., 2020), who used multi-objective RL to train an agent that is able to cause the biggest decrease in long-term rewards while minimising the amount of perturbation. However, in this work, an upper bound of 10% is put on the amount of perturbation instead of minimising it, which simplifies the algorithm into an RL algorithm. This upper bound was also used by (Zheng et al., 2021), who stated that the AC state estimation is not able to detect perturbation up to this threshold. Figure 13 shows the simplified algorithm used to train the perturbation agent.

Algorithm 2: RL-based perturbation agent

Data: $H, K, P, \alpha, \epsilon$
1 Initialize $Q(s, a)$ arbitrarily;
2 **for** $h = i, \ldots, H$ **do**
3 Initialize state s;
4 **while** $k < K$ *and stopping criterion not met* **do**
5 $u \leftarrow Uniform(0, 1)$;
6 **if** $u < \epsilon$ **then**
7 Pick perturbation p randomly from P;
8 **else**
9 $p \leftarrow argmax_{p' \in P} Q(s, p')$;
10 Apply perturbation p to get s^{adv};
11 Let AI agent choose action a^{adv} based on s^{adv};
12 Take action a^{adv} and go to next state s' with reward R;
13 $p' \leftarrow argmax_{\hat{p} \in P} Q(s', \hat{a})$;
14 $Q(s, p) \leftarrow Q(s, p) + \alpha(R + \gamma Q(s', p') - Q(s, p))$;
15 $s \leftarrow s'$
16 **end**
17 **end**

FIGURE 13 – ALGORITHM OF THE RL-BASED PERTURBATION AGENT

Here, *H* is the number of episodes, and *K* is the maximum number of steps in each episode. The set *A* contains all possible actions the perturbation agent can take, which in this case are the possible perturbations, and *α* and *ϵ* are the learning and exploration rates. In each step, the perturbation agent either performs a random perturbation or the best perturbation according to the Q-values (lines 5-9), which are commonly used in RL frameworks and, in this work, represent how well the perturbation agent can lower the performance of the AI agent. Based on this perturbation, the environment will then go to the next state, *s'*, which will come with an immediate reward of *R*. To get an estimate of the long-term value of the perturbation performed, the best perturbation that can be performed in *s'* is chosen in line 11. Afterwards, in line 12, the Q-value for the original state and performed perturbation, *Q(s, a)*, is updated using *R* and the Q-value of the best case perturbation in *s'*.

The perturbation agent can do nothing or perform a perturbation. This perturbation can change one of the values in the observed state and either replace it with zero or a very large number. The agent can also create an adversarial example using the Fast Gradient Sign Method (FGSM) (Goodfellow et al., 2014) to attempt to make the AI agent take a specific action. Since this work is focused on black-box attacks, the gradients used for the FGSM are estimated using Eq. (3.1), which is also used by (Chen et al., 2021). However, where they used these estimated gradients in a projected gradient descent algorithm to create adversarial examples, the FGSM is used in this work. Additionally, Chen et al. applied this perturbation during each step of the digital environment, while we only create the adversarial example if the adversarial agent thinks it will result in the biggest drop-off in performance for the AI agent.

In Eq. (4.4), *L(s)* is the target function for the adversarial example. Since the goal in this work is to make the AI agent perform a certain action, *L(s)* is defined as the value corresponding to the action in the policy of the AI agent. The vector e_i in Eq. (4.4) has a 1 in the i_{th} position and 0 everywhere else.

$$g_i = \frac{\partial L(s)}{\partial s_i} \approx \frac{L(s + 0.01 \cdot e_i) - L(s - 0.01 \cdot e_i)}{0.02} \tag{4.4}$$

After estimating the gradient using this equation for each value in the state *s*, the adversarial example can be computed using Eq. (4.5) as long as the maximum amount of perturbation η is known.

$$s^{adv} = s + \eta \cdot g \tag{4.5}$$

Action perturbing agent: The action perturbing agent can be seen as a layer between the AI agent and the environment. Instead of the AI agent's action being directly executed in the environment, the action-perturbing agent might overwrite or modify the action before executing it in the environment. This mimics the situation where a human operator might confirm or ignore suggestions given by an AI system, as opposed to the RL training setting, where actions suggested by the system might be executed directly in the environment. Thus, the action perturbing agent tests robustness to the shift in conditions between the train conditions (direct execution of AI agent's chosen actions) and test conditions (moderated execution of AI agent's chosen actions).

The action perturbing agent requires a domain-specific choice to be made to define a "simulated operator": a probability is set that the perturbing agent modifies a suggested action at each time step, a number of variables are chosen that is modified in these instances (in case actions consist of choosing more than one variable) and a strategy is set for this simulated operator, that could be based on imitating actions from a logged dataset of human actions if available and might otherwise be a pre-defined heuristic or even completely randomized actions. Then, at every time step, with the chosen probability, the action perturbing agents modify the chosen number of variables according to the set strategy.

Communication Perturbing Agent: The communication perturbing agent is designed similarly to the action perturbing agent. Rather than interfere with the action, it interferes with the messages transmitted between agents in the environment. This perturbation can be designed to simulate various real-world situations arising from 1) variances in communication bandwidth or 2) variances in the communicated messages. Two cases of variance in communication bandwidth exist: 1a) reduction of communication bandwidth by a certain percentage or 1b) no communication bandwidth available. Both cases can occur as the result of system malfunction or adversarial attacks on the system. Perturbations to communication bandwidth test the agents' ability to solve the problem with less or no information from other agents in the system. Perturbations to message content can occur either in the form of 2a) general environmental noise or 2b) incorrect message content. The former case foresees the addition of stochastic noise over communicated messages, whereas the latter considers the case in which specific agents transmit incorrect information. This incorrect information may arise as the result of system malfunctions or be injected in direct adversarial attacks on the system. Evaluating the impact of these perturbations provides important insights into the influence of communicated messages on solution quality and the robustness of the system to variations in message content.

Domain-specific perturbation agents

a) Air Traffic Management

From the qualitative evaluation of the likelihood and impact of perturbations on the AI system during ATM operations, risks related to state/input space perturbations were highlighted as potentially the most disruptive. Consequently, the performance of the AI algorithms for the tasks of sectorisation and path planning will be evaluated against an adversarial agent altering input parameters, like weather information (location of adverse weather cells that need to be avoided), uncertainty in flight entry times in the FIR or algorithm settings (like buffered minimum separation distance).

The simulations of the impact of the perturbation agents will be heuristic, introducing disturbances according to probabilistic rules that are sensibly specified for each type of perturbation. Simulations will be conducted in BlueSky, altering the input to the AI while maintaining the intact state. This will showcase the performance of the AI, used in the planning stage, in actual operations where the real-time situation might evolve in different directions than forecasted.

The perturbation agent for the air traffic management use cases can introduce perturbations to the environment state as described here.

Weather disturbance. This perturbation simulates limitations to the available airspace caused by weather perturbations of varying magnitude and duration. In the BlueSky environment, this is implemented through the activation and deactivation of cells (basic shapes in BlueSky) with time stamps, indicating no-fly zones. These cells will be initialised with a random shape, centred at random locations. The robustness of the AI-generated solutions for the aircraft trajectories and sector plans will be evaluated against the small and large no-fly zones. In addition, given that the AI works in a planning stage, the flow management position supervisor should be able to simulate what-if scenarios, including no-fly zones, selecting or nudging the solution based on their experience and expectations for the weather conditions for the day of operations.

Volcanic ash disturbance. This perturbation follows the same principles as weather-related no-fly zones but is characterized by larger impacted areas and a prolonged duration. Unlike transient weather events, volcanic ash clouds can persist over extended periods, requiring sustained rerouting and airspace restrictions. The performance of the AI-generated solutions will be tested with these longer-term disruptions. Modelling of these no-fly zones in BlueSky will mirror the modelling of bad weather cells but with adjusted parameters. The flow management position supervisor should also explore alternative scenarios, testing different scenarios in airspace closures and evaluating potential contingency measures.

Flight entry time. This perturbation simulates the impact of flights that enter an airspace sector earlier or later than expected, introducing real-time variability. Early arrivals and delayed flights may result in temporary congestion, as well as inefficiencies in planned airspace utilization. AI-based solutions must be robust to these variations, ensuring that sector plans remain balanced in the event of small sector entry perturbations.

The basic logic of the perturbation agent is the same as the one described in the railway domain: at each timestep of the simulation, the adversarial state s^{adv} is updated and used as the AI system's input. This simulates the inaccuracies in the forecast of the input parameters of the AI. The elements of the adversarial state are modified based on a probabilistic process, either introducing a perturbation with probability p_i if none is currently present or removing an existing perturbation with probability q_i.

Weather disturbances (see the algorithm in Figure 14) are introduced as adverse weather cells that appear in the AI's input with a probability $p_{weather}$, which will be empirically determined. These weather cells are initialised as small, randomly shaped regions centred at random locations. Over time, their boundaries evolve probabilistically: the size of the geofence increases gradually after formation and then, after reaching a peak, shrinks until disappearing. The geofence centre may also move dynamically, simulating drifting storm cells. The probability $q_{weather}$, governing their disappearance, follows a Poisson distribution, ensuring that while some disturbances persist, others dissipate relatively quickly, mimicking natural weather evolution.

ATM 1 Weather perturbations

Data: s (state with center and size of cell), $p_{weather}$, $q_{weather}$

Sample $u \sim \text{Uniform}(0, 1)$
$t_{end} \leftarrow -1$
Perturbation flag $\leftarrow$ False
if $u < p_{weather}$ **then**
    Sample $t_{end} \sim \text{Poisson}(q_{weather})$
    Perturbation flag $\leftarrow$ True
end if
if Perturbation flag **and** $t < t_{end}$ **then**
    $s_{\text{size}} \leftarrow \text{updateSize}(s_{\text{size}}, t, t_{end})$
    $s_{\text{center}} \leftarrow \text{updateCenter}(s_{\text{center}}, t)$
    $s \leftarrow (s_{\text{center}}, s_{\text{size}})$
end if
$s \rightarrow$ actions
Go to next timestep $t + 1$

FIGURE 14 – WEATHER PERTURBATIONS IN AIR TRAFFIC MANAGEMENT

Volcanic ash disturbances (see algorithm Figure 15) appear suddenly with a very low probability $p_{volcanic}$, representing the rarity of such events. Unlike transient weather cells, they remain fixed in position once introduced, but their size evolves over time. Initially small, the ash cloud grows gradually before stabilising and eventually shrinking until it disappears. The duration of the volcanic ash disturbance follows a uniform distribution within a range of 6 to 10 hours, with disappearance governed by a Poisson-distributed probability $q_{volcanic}$.

ATM 2 Volcanic perturbations

Data: s (state with size of cell), $p_{volcanic}$, $q_{volcanic}$

Sample $u \sim \text{Uniform}(0, 1)$
$t_{end} \leftarrow -1$
Perturbation flag $\leftarrow$ False
if $u < p_{volcanic}$ **then**
    Sample $t_{end} \sim \text{Poisson}(q_{volcanic})$
    Perturbation flag $\leftarrow$ True
end if
if Perturbation flag **and** $t < t_{end}$ **then**
    $s_{\text{size}} \leftarrow \text{updateSize}(s_{\text{size}}, t, t_{end})$
    $s \leftarrow (s_{\text{size}})$
end if
$s \rightarrow$ actions
Go to next timestep $t + 1$

FIGURE 15 – VOLCANIC PERTURBATIONS IN AIR TRAFFIC MANAGEMENT

Delays in flight entry times are modelled as stochastic variations in the timestamps of aircraft entering the FIR, as shown in Figure 16. These perturbations follow a normal distribution skewed toward delays, reflecting real-world operational trends where late arrivals are more common than early ones. The AI system must account for these variations dynamically, ensuring that sectorization and path-planning adjustments remain effective in mitigating congestion and optimising airspace use.

ATM 3 Aircraft Entry Time Perturbations

Data: t_{entry} (flight plan entry time), μ (mean delay), σ (standard deviation), α (skewness)

Sample $\Delta t \sim \text{SkewNormal}(\mu, \sigma, \alpha)$
$t_{\text{entry}} \leftarrow t_{\text{entry}} + \max(0, \Delta t)$
Update flight trajectory with perturbed entry time t_{entry}
Go to next timestep $t + 1$

FIGURE 16 – AIRCRAFT ENTRY TIME PERTURBATIONS

b) Power Grid

For the use case "Sim2Real, transfer AI-assistant from simulation to real-world operation", see (Bessa et al., 2024) for a complete description of a perturbation agent using the NoisyObservation class in Grid2Op[8] was implemented with the following parameters:

- σ_{line}^{P}, standard deviation used for the active power flow values at the extremity and origin side of each powerline,
- σ_{line}^{Q}, standard deviation used for the reactive power flow values at the extremity and origin side of each powerline,
- σ_{line}^{A}, standard deviation used for the current flow values at the extremity and origin side of each powerline,

[8] See https://grid2op.readthedocs.io/en/latest/user/observation.html#grid2op.Observation.NoisyObservation

- σ_{load}^{P}, standard deviation used for active load values,
- σ_{load}^{Q}, standard deviation used for reactive load values,
- $\sigma_{generator}^{P}$, standard deviation used for active generator values,
- $\sigma_{generator}^{Q}$, standard deviation used for reactive generator values,
- $\sigma_{storage}^{P}$, standard deviation used for active storage values.

The affected attributes are:

- The active power flow values at the extremity (p_{ex}) and origin (p_{or}) side of each powerline that are added with $logNormal(0, \sigma_{line}^{P})$,
- The active power flow values at the extremity (q_{ex}) and origin (q_{or}) side of each powerline that are added with $logNormal(0, \sigma_{line}^{Q})$,
- The current flow values at the extremity (a_{ex}) and origin (a_{or}) side of each powerline that are multiplied by $logNormal(0, \sigma_{line}^{A})$[9],
- The active ($load_p$) and reactive ($load_q$) load values of each consumption, that are multiplied respectively by $logNormal(0, \sigma_{load}^{P})$ and $logNormal(0, \sigma_{load}^{Q})$,
- The active (gen_p) and reactive (gen_q) load values of each consumption, that are multiplied respectively by $logNormal(0, \sigma_{generator}^{P})$ and $logNormal(0, \sigma_{generator}^{Q})$,
- The active storage values storage_power that are added with $logNormal(0, \sigma_{storage}^{P})$.

In each case, the vector's values are multiplied by a vector with lognormal values of same size.

A second perturbation agent for this environment is the **random perturbation agent,** which introduces random perturbations to simulate potential failures in the SCADA or state estimation system. In other words, it replicates natural adversarial disruptions to the system, such as missing measurements. The random perturbation agent can act in complement to the noisy observation defined previously:

- noisy observation can model the intrinsic noise due to measurement (e.g. sampling), which continuously applies to all variables of the environment (even in absence of particular issues),
- on the other hand, the random perturbation agent can model a specific measurement issue, which will happen in addition to the intrinsic noise.

At each step in the environment – corresponding to each instance when the dispatcher monitors the system – the agent can inject a new perturbation, as illustrated in Figure 17.

[9] This implies that the load of the line (defined as the observed current flow divided by the thermal limit) is also multiplied by the same factor.

Algorithm 3: Random perturbation agent rules

Data: $p, PP, s^{gen}, s^{load}, s^{flow}, \sigma^{gen}, \sigma^{load}, \sigma^{flow}$

1 Apply perturbations in PP to s^{gen}, s^{load} and s^{flow};
2 $u \leftarrow Uniform(0, 1)$;
3 **if** $u < p$ **then**
4 Randomly choose value to perturb s_i^m where $m \in \{gen, load, flow\}$ and $i \in \{1, \ldots, |s^m|\}$;
5 $u \leftarrow Unif(0, 1)$;
6 **if** $u < 0.2$ **then**
7 $s_i^m \leftarrow 0$;
8 **else**
9 $r \leftarrow logNormal(0, \sigma^m)$;
10 $s_i^m \leftarrow s_i^m \times r$;
11 $k \leftarrow Geometric(\frac{1}{6})$;
12 Add new perturbation to PP for k steps;
13 Update remaining steps for perturbations in PP;

FIGURE 17 – ALGORITHM OF THE RANDOM PERTURBATION AGENT FOR THE POWER GRID

As an input, the random perturbation agent needs the probability of introducing a new perturbation *p*, the set of previous perturbations that should still be applied *PP*, the actual data on generation (s^{gen}), load (s^{load}) and flow over power lines (s^{flow}) and the standard deviation of the introduced perturbation for each data group σ^{gen}, σ^{load}, and σ^{flow}. In lines 2 and 3 of the algorithm, a new perturbation is introduced during a step with probability *p*. If this is the case, then one specific value is chosen to be perturbed. The perturbation can either mimic a complete failure to measure a value (line 7) or an incorrect measurement (lines 9-10). In the case of an incorrect measurement, the lognormal distribution is chosen to ensure that the sign of the value remains the same and because we assume that the measurement errors approximately follow a normal distribution. This perturbation is then applied in the next *k* steps by adding it to *PP*, with an average length of 6 steps, equivalent to 30 minutes. The geometric distribution is chosen for *k* because of the memoryless property of the distribution.

Finally, it is important to mention that Grid2Op already has an opponent to test the robustness under contingency scenarios due to extreme weather of cyber-attacks, which uses different heuristics to define attacks on the power lines (Omnes et al., 2021).

c) Railway

For the generation of adversarial datasets for the railway use cases, a similar approach with perturbation agents as for the power grid cases will be followed: a perturbation agent can alter the input space of the AI system, and such an agent can be employed during training and testing. For the railway cases, the Flatland environment will be used, a railway simulation environment capturing the relevant aspects for the railway use cases. The state of the Flatland environment, which also represents the input space to the AI system, notably includes the position, speed, and schedule of the trains and the track layout of the railway network.

The perturbation agent considered for the railway cases is based on heuristics and introduces stochastic but interpretable disturbances into the railway system state. The perturbation agent does

not actually alter the state of the environment; however, it will preserve perturbations over time to mirror real-world situations, like the malfunction of a track occupancy sensor. Therefore, the perturbation agent acts effectively as a transformation layer between the environment state provided by the Flatland environment at each timestep and the input space for the AI system, which, for example, can be used to build the observations for RL models.

The perturbation agent for the railway cases can introduce four types of alterations to the environment state described in the following.

Track availability. The track availability perturbations simulate incomplete information about the availability of a track segment. Such incomplete information could be due to defective sensors (e.g., incorrectly reporting that a switch is in the wrong position or that the information is missing completely) or loss of information during transmissions, among other things. In the Flatland environment, the track availability is reflected by changes in the railway network topology. An unavailable track or a switch that reports the wrong position is modelled by removing the corresponding transition from the transition map that is part of the environment state[10].

Track occupancy. The track occupancy is similar to the track availability in its effect, but the source of the perturbation is different. The track occupancy reports if a train is currently on a track. A perturbation can reflect a broken sensor detecting if a train is currently on a track segment or not. This is a binary perturbation, i.e., a broken sensor would result in either no detection of trains at all for a specific segment or a permanent report of a train occupying the track. In the first case, a train position would not be reported while the train is actually on a track segment, i.e., the train position would be removed from the state during the time steps the train would be reported on that segment. In the second case, the sensor would permanently report that a train is on the track segment, effectively introducing a virtual train into the system. In the Flatland environment, the track occupancy is altered by either removing the position of a train from the environment state while it is on the perturbed segment (first case) or introducing a non-moving virtual train with its position fixed on the perturbed track segment for an extended time (second case).

Train location. The train location perturbation reflects problems with reporting a train's location through a system other than track occupancy. For example, the transmission of a train's location could experience delay such that the train's location at a specific point in time is behind the actual location of the train at that time. In this case, a train's location, i.e., the position on the Flatland environment grid space, is shifted in the direction where the train is coming from. A second example is that the location of the train can only be determined with low accuracy, such that the location could be reported in front of or behind the actual location on the track. In this case, a train's position in the environment space is shifted to any connected track segment. The train location case is closely related to the track occupancy. While the location of the track influences the location of a train in the track occupancy case, a train's location is altered based on its own location in the case of the train location.

[10] see https://flatland-association.github.io/flatland-book/ for the documentation of the environment, including the transition maps and state

Train schedule. The train schedule perturbations are very different from the other three perturbation mechanisms. This case reflects wrong data about a train's schedule, potentially affecting both the stops of a train and the arrival and departure times at the respective stations. Such corrupt data could be due to human error or to malfunctions in the system that manages the schedule. In the Flatland environment, such a perturbation is modelled by changing a train's reported target stations and the latest arrival and earliest departure times for the stations.

The perturbation agent introduces disturbances to the input space through the four mechanisms described above, as shown in Figure 18.

Algorithm: Introduction of perturbations into state space

Data: $s, \mathcal{P}_k, p_k, q_k, \mathfrak{p}_k, t$

1 **for** i **in** s **do**
2 **for** k **in** perturbation type allowed at s_i **do**
3 **if** s_i has no perturbation k **then**
4 with probability p_k **do**
5 $\mathcal{P}_{k,i} \leftarrow \mathfrak{p}_k$;
6 $t_{k,i} \leftarrow t$;
7 **else**
8 with probability $q_k(t - t_{k,i})$ **do**
9 $\mathcal{P}_{k,i} \leftarrow$ None;
10 **end**
11 **end**
12 **end**
13 $s^{adv} \leftarrow s \oplus \sum_k \mathcal{P}_k$;
14 $s^{adv} \rightarrow$ actions;
15 Go to next timestep $t + 1$;

FIGURE 18 – ALGORITHM OF THE PERTURBATION AGENT FOR THE RAILWAY NETWORK

At each timestep of the environment, the adversarial state s^{adv} consists of the environment state s superimposed with the perturbation matrix for each possible perturbation. The adversarial state is then used as input space for the AI system instead of s (line 14). The perturbation matrix contains all perturbations k, which for each element are added with probability p_k if there is currently no perturbation or otherwise removed with probability q_k. The probabilities p_k follow a uniform distribution while the probabilities q_k follow a Poisson distribution, creating a natural life cycle for the perturbations, mimicking maintenance behaviour, with many perturbations being resolved rather quickly while others persist for an extended period. To keep track of this progress, the timestep is saved for each element and perturbation type. While the method to add a perturbation depends on the mechanism described above, the general stochastic principle when such a method is called to set the value of an element in the perturbation matrix remains the same for all four mechanisms. However, the probabilities for adding and removing perturbations are different for the four types of possible alterations described above. The type of alteration is fully determined by the index i of the states s_i and $s^{adv}{}_i$, and the corresponding probabilities p_k and q_k are defined by railway domain experts.

4.3.3 EVALUATION METRICS

The robustness and resilience of the system will be quantified using the perturbation agents outlined in Section 4.3.2. These agents introduce environmental perturbations impacting AI system performance, replicating natural and intentional scenarios, including subtle, imperceptible disturbances.

4.3.3.1 ROBUSTNESS

This subsection describes metrics specifically defined in WP4 to measure different properties of robustness in AI systems that are not domain-specific.

The first metrics (**KPI-RS-058** and **KPI-DF-069** in Annex 1) measure how overall system performance changes due to perturbations. Such perturbations can originate from different sources: through unexpected changes in the environments (e.g., natural malfunctions or intentional attacks) or through operator actions, since the final decisions from the operator could bring the system to states that are not encountered frequently in simulations without an operator present. **Robustness to operator input** KPI-RS-058 focuses on the last type of perturbation and considers the system's performance when not all AI agent suggestions are executed by a simulated operator. In the fully autonomous case, the human agent has the ability to provide high-level directives, which are a type of perturbation to the system that would be covered by this metric. **Drop-off in reward** KPI-DF-069, on the other hand, focuses on perturbations of the environment and measures the difference in total rewards between the unperturbed and perturbed AI systems. The metric measures if the AI system can perform at the same level when introducing perturbations and can be calculated using the methodology in Annex 1.

Another factor that can be used to determine the robustness of an AI system is the range of change in the output of the system when perturbations are introduced. In this case, the output would be the action recommended by the AI agent, and the change in this action is measured using two approaches. The first approach is to assess whether a particular decision holds for input variation (e.g., noise, missing data) in the same context by counting the number of times the decision the AI system takes with perturbations is different compared to the decision the system takes without any perturbations, as shown in **Frequency changed output AI agent KPI-FF-070** in Annex 1. The second approach to measure the range of change in the output is to not only look into when the action is changed but also to look at how similar or different the new action with perturbation is to the original one. To do this, a similarity score is assigned to each pair of actions, as described in **Severity of changed output AI agent KPI-SF-071** in Annex 1. These metrics enable the verification of the stability property as outlined in ISO/IEC 24029-2. This property evaluates whether the system's output remains consistent despite variations in the input and whether its performance is maintained under such conditions. In this case, stability is measured compared to the expected output of a pre-trained AI agent without any perturbation.

Another metric is the **number of steps in each episode before a critical state is reached** (e.g., power grid failure occurs due to a cascading event), as described in **KPI-SF-072** in Annex 1. This metric aligns

with the reachability property defined in ISO/IEC 24029-2, as it evaluates whether the AI agent can successfully prevent reaching a critical state under perturbations.

Even if the perturbations are not able to decrease the obtained reward or cause a grid failure, they might cause the AI system to take unnecessary actions, which usually comes at a cost. Because of this, **reward per action** is another metric used to measure performance – see **KPI-RF-078** in Annex 1.

Additionally, examining data points that act as weak spots in the system – those capable of significantly altering the system's output with minimal changes – can provide valuable insights. This can be measured by looking at the proportion of times a perturbation led to a change in action by the AI system, which can be computed with **Vulnerability to perturbation KPI-VF-073** in Annex 1.

These metrics are intended to complement the ones defined in WP1 and further refined in Section 3 - Technical performance and scalability (technical performance), where the technical performance of the AI-based system is compared under conditions both with and without perturbations. For instance, for the power grid, carbon intensity, network utilisation, and topological action complexity can be calculated with and without the perturbation agents. This approach also enables the evaluation of the impact of these perturbations on the entire system (beyond just the performance of the AI system) while considering metrics of social significance, such as train delays and carbon emissions.

In addition to actions and their subsequent rewards, some agents in this project are designed to output a corresponding explanation for the action taken. These explainable AI (XAI) methods must too be evaluated for their robustness in the face of perturbations. Evaluating XAI methods remains a critical challenge in research, with no consensus on the best evaluation approach. However, explanation methods can be evaluated using computational metrics, which do not require human participants and allow for more systematic comparisons. We focus on two metrics to assess the robustness of explanation methods.

Explainability Faithfulness (**KPI-EF-087** in Annex 1) measures how well explanations align with the model's predictive behaviour. The underlying assumption is that more important features should have a greater influence on predictions. This KPI uses the Faithfulness Estimate Metric (Alvarez-Melis, et al., 2018) to evaluate whether explanation scores truly reflect the importance of features in a model's decision-making process. It does this by systematically removing or altering features and observing how the model's predictions change. The assumption is that if an explanation method assigns high importance to a feature, removing that feature should significantly impact the model's output. Evaluating faithfulness is performed at run-time with the baseline scenario, after determining the most important features in an RL step. **Explainability Robustness** (**KPI-EF-086** in Annex 1) evaluates the stability of explanations against small input perturbations, assuming the model output remains relatively unchanged. This is based on the Average Sensitivity Metric (Yeh, et al., 2019), which quantifies how stable an explanation is to small perturbations in the input. XAI robustness can either be evaluated at run-time with the baseline scenario by applying small perturbations or using one of the generic perturbation agents defined in Section 4.3.2 (on the subset of scenarios in which the agent's action does not change).

4.3.3.2 RESILIENCE

The quantification of resilience is strongly related to the magnitude and duration of reward function performance degradation compared to an unperturbed system in the same context. The first metric **KPI-AF-074** described in Annex 1, is the **area between the reward curves of the unperturbed and perturbed AI system** from the episode where the perturbations are introduced, as depicted in Figure 19.

The next metric, **KPI-DF-075** described in Annex 1, measures how quickly the AI system can **adapt to the introduction of perturbations** by counting how many episodes the degradation and restorative stages consist of. An example is also depicted in Figure 19.

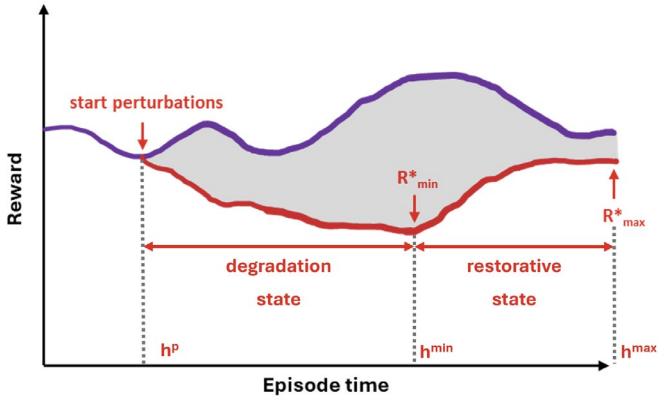


FIGURE 19 – DEGRADATION AND RESTORATIVE STATE DURING THE TESTING OF THE AI SYSTEM

Besides these metrics, it is valuable to examine the extent of performance deterioration and assess whether the system can recover to its original performance level without perturbations. To do this, the minimum reward in the degradation state and maximum reward in the restorative state can be computed with **Restorative time KPI-RF-076** described in Annex 1.

Finally, the **similarity between the state of the environment with unperturbed and perturbed AI systems over time** is used as a metric – **KPI-SF-077** described in Annex 1. It measures how drastically the actual state of the environment is affected after an action of the AI system is changed by perturbations and whether the AI system is able to revert these changes. Calculating this similarity at each step enables the identification of both degradation and recovery states in relation to the unperturbed state, like the approach used for the reward.

4.3.3.3 RELIABILITY

The factors that can affect AI reliability can fall into three levels: system, data, and model (i.e., algorithm). Figure 20 shows the Venn diagram for the three factors that affect AI reliability, which will be further described.

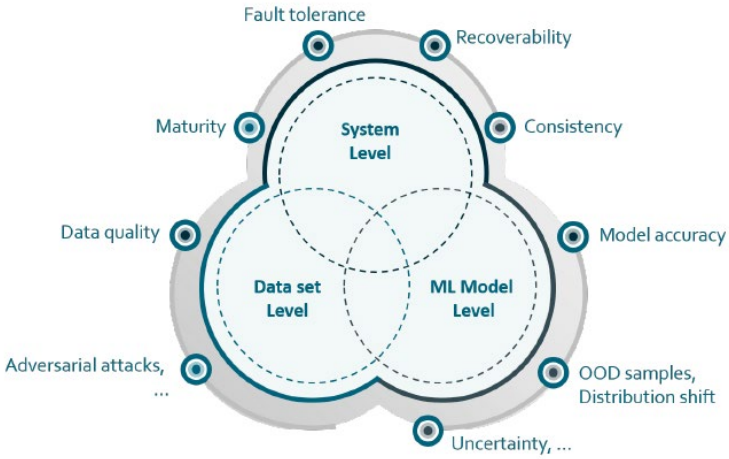


FIGURE 20 – FACTORS THAT CAN AFFECT AI RELIABILITY CAN FLL INTO THREE LEVELS

1. **System Reliability:** Metrics such as failure rate, error rate, MTBF (mean time between failures), and MTTF (mean time to failure) are essential for characterising the reliability of AI systems. MTBF measures the average time between failures, while MTTF indicates the time until the first failure, with higher MTBF signalling greater reliability. Reliability can be empirically assessed using the probability of failure (Pf), calculated as the ratio of failed items to tested items, with reliability $R = 1 - Pf$. However, traditional reliability metrics like MTTF are less applicable to data-driven AI systems, necessitating tailored criteria. Additional metrics like fault correction and test coverage, proposed by ISO/IEC 25023 (2015), are more relevant during the design and testing phases but less useful for evaluating system reliability during operation.
2. **AI model Reliability:** Reliability issues in AI models can stem from two key factors: poor performance in tasks that are typically easy for humans and vulnerabilities that cause malfunctions in specific conditions, whether naturally occurring or adversarially induced. A common challenge for AI models is handling out-of-distribution (OOD) inputs, which are data samples differing from the training distribution. Reliable models must not only perform well on in-distribution (ID) data but also effectively detect and manage OOD inputs by labelling them as “unknown”. Failure to recognise OOD samples can lead to incorrect decisions and errors, making OOD detection critical for enhancing AI system reliability.
3. **Data reliability.** Many decision errors are caused by distribution shifts where the AI inference differs from the AI training. The distribution shift between datasets is defined as the distance between datasets according to the defined properties of these datasets. Adversarial attacks and data quality issues significantly impact the reliability of AI-based systems by causing inaccuracies in predictions and software failures. Adversarial attacks involve small data perturbations to deliberately mislead models, while noisy or imbalanced data can degrade model accuracy, particularly in classification tasks. Data quality, including biases and imbalances, directly affects an AI algorithm’s performance. Reliability in AI systems depends on the robustness of algorithms against such attacks and the quality of the training data. The concept of data reliability, as defined by EUROCAE WG114 – SAE G34 (2021), involves

confidence in data sources and measurements of reliability, such as internal, relative, and absolute reliability, based on the proportion of correct data items in a dataset.

The data reliability issues that are mentioned here are captured by the robustness metrics introduced earlier in this section. The focus of the reliability metrics is more on handling the out-of-distribution data, which may be related to domain shifts, which are the subjects of the next part.

Domain shift in RL and related metrics

In machine learning, model performance and reliability heavily depend on the quality and consistency of input data. However, real-world data often changes over time, a challenge known as **Domain Shift** or **Data Drift**, where the statistical properties of input or target data shift, potentially degrading model performance. Detecting data drift is crucial to maintaining model accuracy, as failure to do so can lead to reduced efficiency and risks in decision-making. Data drift can stem from evolving user behaviour, changes in data sources, or external factors, and it manifests in various types. Effective drift detection enables timely corrective actions, preserving the reliability and effectiveness of predictive models. A robust AI system should be able to adapt and update its model or algorithms to account for concept drift and maintain its accuracy. Mathematically, the data drift is a change in the joint distribution:

$$\boldsymbol{P}(X_{ref}, Y_{ref}) \neq \boldsymbol{P}(X_{target}, Y_{target}), \tag{4.6}$$

where $\boldsymbol{P}(X_{ref}, Y_{ref})$ is the reference joint distribution and $\boldsymbol{P}(X_{target}, Y_{target})$ is the target joint distribution. In practice, these two distributions can refer to the training and the testing distributions respectively. Note that X represents the features and Y are their corresponding targets.

This equation suggests that data drift can occur either because of a change in the distribution of features or targets or because of their conditional distribution. There are different concepts related to the problem of dealing with changes from the training domain to the testing domain (Haider, et al., 2021):

- **Distributional shift and OOD:** *Distributional shift* refers to changes in data distributions. When only the input distribution changes but the output distribution remains the same, it is known as a covariate shift and could be formulated as $\boldsymbol{P}(X_{ref}) \neq \boldsymbol{P}(X_{target})$. If the testing distribution differs from the training distribution, machine learning systems may perform poorly and incorrectly assume their performance is satisfactory (Amodei, et al., 2016). When the relationship between the input features and the target variable changes, it is called *Concept drift* and could be formulated as $\boldsymbol{P}(Y_{ref}|X_{ref}) \neq \boldsymbol{P}(Y_{target}|X_{target})$. This means that the function that maps the feature space to the target space changed. OOD refers to data outside the training distribution. Formally, if P_X and Q_X are distinct data distributions on the input space X, a model trained on data from P_X considers those samples as in-distribution, while samples from Q_X belong to the OOD domain. In RL, OOD could be defined as every state-action tuple not experienced during training (Sedlmeier, et al., 2019). As an example, in the power grid use case of the AI4REALNET project, the models could be trained on a set of specific topologies and evaluated on some slightly different topologies at the test time.

- **Domain shift success rate drop**, as described in **KPI-DF-057** in Annex 1: The percentage decrease in the success rate of the agent in the shifted domain compared to the original domain. It indicates how often the agent fails to complete the task in the new domain compared to the original:

$$Success\ Rate\ Drop = \frac{Success\ Rate_{original} - Success\ Rate_{shifted}}{Success\ Rate_{original}} \times 100 \quad (4.7)$$

- **Out-of-domain detection accuracy**, as described in **KPI-DF-054** in Annex 1: The accuracy with which the agent can detect whether it is operating in a domain that is different from the one it was trained on. It is useful for systems that need to switch strategies or request human intervention when a domain shift is detected. A recent paper (Nasvytis, et al., 2024) introduces various approaches for the detection of OOD in RL.

- **Transfer learning, domain adaptation and domain randomisation:** The objective of *transfer earning* is to learn a task T_B that belongs to a target domain B, by using experience (prior knowledge) from some source task T_A from source domain A, i.e. to transfer knowledge from A to B. *Domain adaptation (DA)* is a sub-field of transfer learning. DA can be defined as the capability to deploy a model trained in one or more source domains into a different target domain. While DA encompasses cases where the source and target domains have the same feature space, TL includes cases where the feature space of the target domain differs from the source domain. *Domain randomisation* involves randomly altering the source domain parameters to train a more generalisable model. Studies ((Sadeghi, et al., 2017), (Tobin, et al., 2017)), demonstrate that randomising simulator settings can help train policies that generalise to real-world scenarios without needing highly realistic simulations. (Rajeswaran, et al., 2017) show that randomising system dynamics can create more robust policies that generalise across various target domains, including unmodeled effects in the training distribution.
  - **Adaptation time and performance drop**, as described in **KPI-DF-052** in Annex 1: The time or number of episodes required for the agent to regain a specific level of performance in the shifted domain after the domain shift has occurred. It can be used to evaluate how quickly an agent can adapt to new environmental conditions. However, the adaptation time should also consider the performance drop in its computation, as a high-performance drop after the adaptation could not be tolerated.

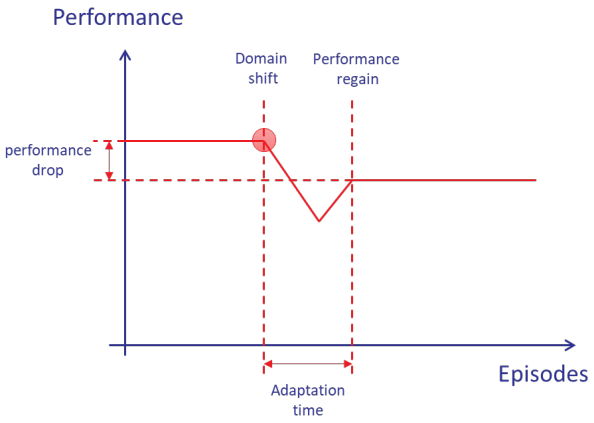


FIGURE 21 – SCHEMATIC REPRESENTATION OF PERFORMANCE DROP AND ADAPTATION TIME

- **Forgetting rate**, as described in **KPI-DF-090** in Annex 1: The rate at which an agent forgets its performance in the original domain after being exposed to a shifted domain. It helps to measure the extent to which learning in the new domain negatively impacts the agent's ability to perform in the original domain.

- **Novelty detection and intrinsic motivation:** Novelty is a mechanism humans use to explore and learn, supported by neuroscience studies showing it activates the brain's reward region in both humans and animals, playing a key role in RL (Houillon, et al., 2013). In RL, however, identifying undesired operations is more closely related to distributional shift and OOD detection, while novelty is commonly used as a mechanism for exploration purposes. Seeking novelty can lead to unexpected, sometimes harmful outcomes. While learning from negative experiences is useful, in safety-critical applications, certain dangerous states must be avoided. Therefore, safe exploration methods are essential to prevent breaching safety constraints.
- **Robust RL:** Robustness in machine learning is crucial because models are expected to perform well despite small input deviations. However, adversarial attacks reveal that neural networks can be vulnerable to minor, specially crafted input changes, leading to drastic prediction errors. This vulnerability also affects RL models. (Nilim, et al., 2003) associate RL robustness with uncertainties in the transition matrix of Markov Decision Processes (MDPs), while (Pinto, et al., 2017) highlight the need for RL models to be robust against model initialisation and errors.
  - **Robustness to domain parameters**, as described in **KPI-DF-056** in Annex 1: The sensitivity of the agent's performance (e.g., reward) to changes in specific domain parameters (e.g., generator type including renewables in power grid domain). It helps to identify which environmental factors most affect the agent's performance.
  - **Policy robustness**, as described in **KPI-DF-055** in Annex 1: a ratio of the performance in the shifted domain to the performance in the original domain. A score close to 1 indicates high robustness, while a lower score indicates reduced performance due to the domain shift. It can assess the generalisation of a policy learned in a simulated environment when applied to a real-world scenario.

$$Policy\ Robustness = \frac{R_{shifted}}{R_{original}} \tag{4.8}$$

- **Operational design domain (ODD):** ODD describes the specific conditions under which automated driving systems are designed to operate safely. It includes details like roadway types, geographic areas, and speed ranges. Documenting how the system should respond when it moves outside its defined ODD is also a critical aspect. This could be mapped easily to the AI4REALNET use cases. For example, in the railway domain, the different constraints on the train's speed, the railway connections and types could define the conditions under which the system could operate safely.
  Evaluation metric on **generalisation gap**, as described in **KPI-DF-053** in Annex 1: The absolute difference between the performance (e.g., rewards) in the training domain and the shifted domain. This metric quantifies the extent of performance loss due to domain shift:

$$Generalization\ Gap = \left|R_{source\ domain} - R_{target\ domain}\right| \tag{4.9}$$

- **Data drift and temporality:** Data drift is inherently connected to temporality since it refers to the changes in the data distributions over time. Temporality plays a crucial role in understanding and managing data drift, as different types of drift have different patterns and might require a different detection solution. The four predominant data drift patterns are widely recognised as follows:
  - Sudden data drift happens when there is an abrupt and significant change in the data distribution. This drift pattern can occur due to unexpected events, abrupt system failures, or sudden shifts in user behaviour.
  - Gradual data drift refers to a slow and continuous change in the data distribution over time. This type of drift often occurs due to gradual shifts in the underlying processes or external factors influencing the data.
  - Incremental data drift occurs when small, incremental changes accumulate over time. These changes are small and may require some time to become noticeable, but they can have a substantial impact when they accumulate over an extended period.
  - Recurring data drift refers to a cyclical or periodic pattern of drift that repeats at regular intervals. This can be due to seasonal variations, economic cycles or any other periodic influence.

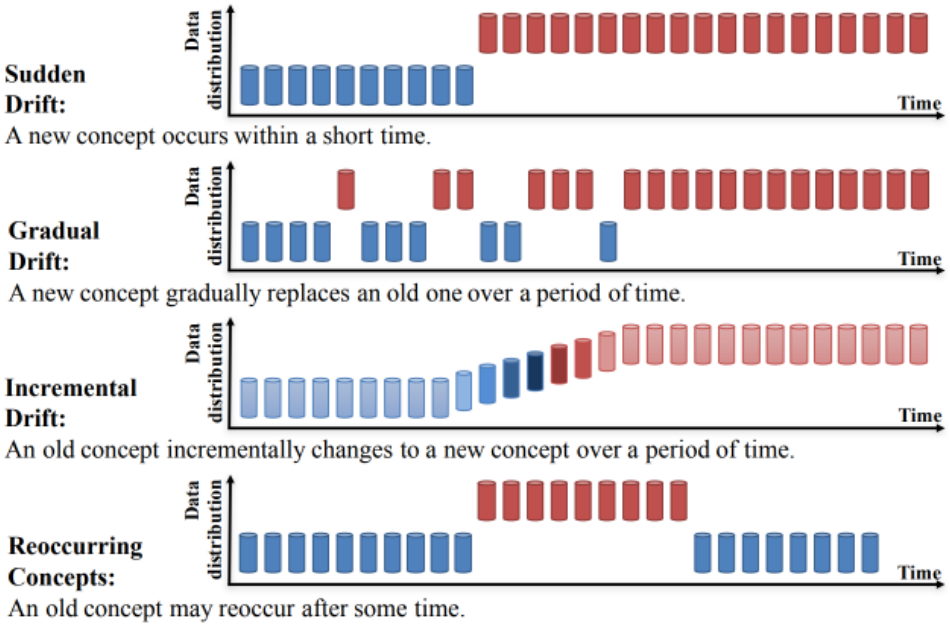


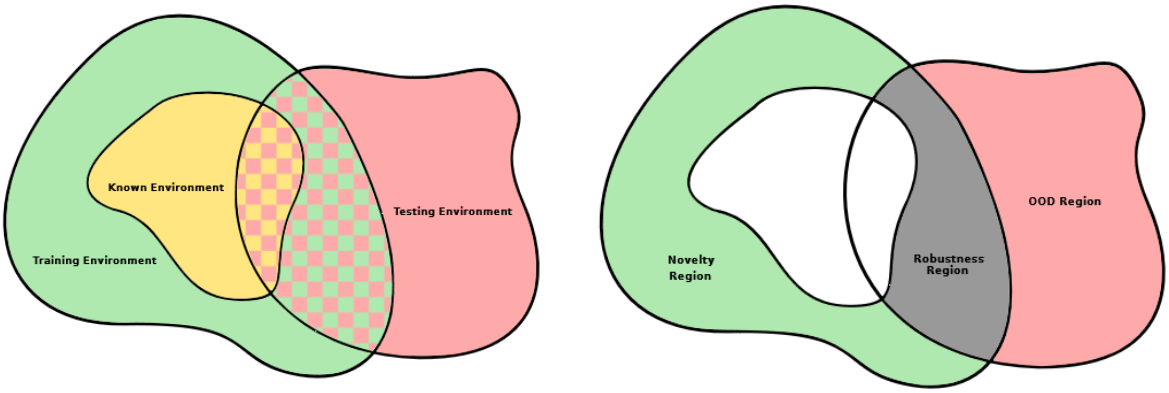
FIGURE 22 – DATA DRIFT PATTERNS AS DESCRIBED (LU, ET AL., 2020)

Considering the different temporal drift patterns is crucial is essential for implementing effective monitoring and adaptation strategies to ensure the continued accuracy and reliability of machine learning models. Each pattern brings its own challenges and requires an adapted strategy to handle them effectively.

These concepts are relevant when dealing with uncertainties and changes in the environment. The distinction between them is subtle and bringing them down to RL problems is not straightforward. Figure 23 (a) illustrates the entire problem domain, representing the distribution of environments the agent may encounter during both training and testing. The **Training Environment** encompasses all scenarios the agent might face during training, also referred to as the in-distribution domain. Using the ODD framework to design this region helps establish safe boundaries that the agent must adhere to. The section of the training domain that has been thoroughly explored is termed the **Known Environment**. Finally, the **Testing Environment** covers the full range of environments the agent could encounter post-deployment. As shown in the figure, there may be overlaps between these domains. The ideal scenario occurs when the testing domain is fully contained within the training region, allowing the agent to explore the entire domain during training—a goal that is often challenging for complex problems.

In addressing domain shifts in RL mapping the regions in Figure 23 (a) to the previously outlined concepts aids in their characterisation, as shown in Figure 23 (b). This mapping is not exhaustive and does not aim to encompass the entire theory of domain shifts. Instead, it suggests a relevant mapping to RL problems that experience environmental changes. It is important to note that unsafe states can still arise within the **known environment** (white region) of a system. The area within the training domain that remains unexplored by the agent is identified as the **Novelty Region**, representing the part that can be explored during training. Post-deployment, the agent should maintain robustness to novel states within the same distribution as the training examples, defined here as the **Robustness**

Region. The final area is the **Out-of-Distribution (OOD) Region**, which consists of samples from distributions that are different from the training distribution.



a) distinction between training, known and testing domains

b) A possible mapping for the domain regions into ML approaches that deal with domain shifts

FIGURE 23 – VISUALIZING THE PROBLEM. FOR RL PROBLEMS, THE REGIONS OF INTEREST CAN BE RELATED TO NOVELTY, ROBUSTNESS AND OOD PROBLEMS (HAIDER, ET AL., 2021)

A limitation of this approach is the difficulty in clearly defining boundaries between these regions, particularly since identifying environmental changes that lead to scenarios outside the training domain is not straightforward. Compounding this challenge is the lack of consensus on how certain concepts should be applied within the RL context. As a result, varying interpretations may emerge depending on how these ideas are adapted from broader machine learning problems to the RL framework. Haider et al. suggest an approach based on Markov decision process (MDP) decomposition. That is, any disturbance is subdivided into the individual components that build up the MDP. This decomposition should come naturally since MDPs are used to describe RL problems in the first place.

To explain this decomposition, we consider two tasks, M_A (training scenario A) and M_B (testing scenario B), which can differ in the following aspects, individually or in combination:

- **S (State-space):** Changes such as enlargements, reductions, or modifications to the set of possible states.
- **A (Action-space):** Variations in available actions, including additional or fewer actions or adjustments to the ranges of continuous actions.
- **P (Transition Dynamics):** Alterations in transition probabilities for the same state-action pairs.
- **R (Reward Function):** Modifications to the reward function that encourage different behaviours.
- μ_0 **(Initial states):** Differences in the distributions of initial states.

Although this may initially seem simple, the method's practical benefits quickly emerge. Training RL agents to meet safety constraints is complex, as modelling every safety-critical situation is infeasible due to real-world complexity. Assumptions about the relationship between training and testing environments are necessary for ensuring safety. Traditional formulations like distributional shift or

robustness offer limited insights for safety, whereas MDP decomposition provides a clearer, component-wise approach to identifying and addressing problems.

Instantiating on Power Grid assistant use case

In this use case, the AI-based RL system is designed to assist human operators in decision-making hours before a critical event occurs (e.g., a power line disconnection or grid anomaly). The RL agents are trained on a source domain consisting of a limited set of observations (state-action pairs), representing only a small subset of all possible scenarios in a power grid. For instance, they might observe a restricted range of topological actions available for each substation. Despite being trained on scarce data, these agents must generalise their learned transition dynamics to assist human operators in situations that either slightly deviate from the training domain (previously unexplored conditions) or fall entirely outside it (out-of-distribution scenarios). An example of such a challenge would be handling a higher-order power line disconnection (N-2 contingencies) caused by environmental factors or adversarial attacks, whereas training was limited to N-1 contingencies. A similar situation could be observed when human operators interact with AI-based assistants. As an example, the decisions made by multiple human operators may be very different and maybe antagonistic with respect to the same context. That could reduce the reliability of the agent for suggestions of relevant assistance.

A similar challenge arises when human operators interact with an AI-based assistant. For instance, different operators may make vastly different—even contradictory—decisions when faced with the same contextual scenario. This variability can undermine the reliability of the AI system in providing relevant and consistent assistance, as it struggles to adapt to diverse human decision-making patterns.

Identifying the nature of disturbances in the target domain is crucial for implementing specific safety countermeasures and ensuring robustness. However, categorising even simple examples under definitions such as distributional shift, novelty detection, out-of-distribution scenarios, or robustness can be complex and may provide only limited insights for building a strong safety argument. As discussed in the previous section, decomposing the MDP allows for a structured analysis of these challenges by isolating issues within individual MDP components. This decomposition is particularly valuable when reasoning about safety, as it enables a systematic identification of potential hazards within the power grid domain, ensuring a more targeted and effective approach to risk mitigation. As such, it provides a decomposition of potential hazards in the power grid domain into components of MDP.

	S	A	P	R	μ_0
New topological actions (creation of new substations)	✓	✓	✓		
Dependency on external systems (weather forecast error, noise in SCADA sensors)	✓				
Planned and unexpected outage events	✓		✓		

	S	A	P	R	μ_0
Introducing more renewable energies (new goal - maximise their usage)	✓			✓	
Perturbed agent (adversarial attack)		✓	✓		
Unusual starting point (not relevant for the power grid maybe, but it can make sense for the railway, for example)					✓

TABLE 5 – DECOMPOSITION OF POTENTIAL HAZARDS IN THE POWER GRID DOMAIN INTO COMPONENTS OF MDP

5. SOCIAL-TECHNICAL DECISION QUALITY

This section details contribution of Task 4.3 to the evaluation, using a common template to define evaluation protocols. It provides an overview of evaluation protocols (including individual evaluation metrics and their clusters corresponding to evaluation objectives) that focus on the human-centered aspects of the AI solutions proposed in the project, considering not only the technical performance of the models and software, but also the overall performance of AI-human collaborations and attitudes of human operators towards such AI solutions.

The Task 4.3 is planned to start during M18, i.e., during the final month of preparation of this report, and thus the considerations described below rely on analyses of the previously defined conceptual project framework and preliminary proposed metrics defined for use cases in deliverable D1.1 (AI4REALNET framework and use cases), as well as initial discussions with the partners involved in the respective task. The described protocols provide an outline of the evaluations to be performed until M42, however, specific experimental designs for some of such evaluation studies cannot yet be defined in detail at this point and will be subject to active research work conducted by the academic partners involved in Task 4.3.

5.1 CONTEXT

While ensuring high levels of performance from the technical standpoint is a crucial foundation for AI solutions (see §3 - Technical performance and scalability and §4 - Safety and robustness), it is equally important to consider the intended context of real-world applications of such AI solutions to be developed within the scope of the AI4REALNET project.

AI-based digital assistance and decision support systems are intended to be employed by **human operators** across several domains. The initial **introduction, learning, trust building, and usability** assessment with human operators is important for ensuring successful long-term adoption of the proposed solutions; one of the expected scientific impacts of AI4REALNET project is to “Increase AI trust” (see the Project Proposal, Section 2.3). Therefore, several groups of human and human-AI teaming concerns are evaluated as part of Task 4.3 in this project. Overall, the AI4REALNET project aims to achieve “>80% of acceptance rate by human operators” (project outcomes described in the Project Proposal, Section 2.3), further detailed as the overall project KPI-ET-7 "% of acceptance of human operators regarding AI4REALNET solutions: 80% (surveys, interviews)" (see the Project Proposal, Section 2.1.1).

In the rest of this section, the overarching evaluation objectives, the respective evaluation metrics and their clusters, and considerations for evaluation protocols and analyses of the respective findings are discussed.

5.2 EVALUATION PROTOCOLS

5.2.1 METHODOLOGY

While the performance of machine learning algorithms is typically evaluated using metrics calculated for a specific test run (or a group of test runs), human operator-related concerns require further considerations and experimental methods studied in the disciplines and fields such as applied psychology, cognitive science, human factors (including human-computer interaction), and human-centered AI, often in relation to the respective design guidelines (see (Amershi, et al., 2019), (Liao, et al., 2020) and (Shneiderman, 2020)). Evaluation methodologies, measurements/metrics, experimental designs, and further considerations relevant to applied psychology and HCI are further discussed in detail by (MacGrath, 1995), (Frøkjær, et al., 2020), and (Kosch, et al., 2023). The respective concerns specifically relevant to human-centered machine learning and explainable AI are discussed by (Sperrle, et al., 2021), (Hoffman, et al., 2023), (Silva, et al., 2022), and (Nauta, et al., 2023).

The following methods are particularly relevant to socio-technical decision quality evaluation in the scope of Task 4.3 of AI4REALNET:

- Task-based user studies focusing on usability of proposed solutions (typically measuring errors and time taken by participants to complete tasks),
- Psychophysiological measurements (e.g cardiovascular activity, eye tracking, etc.),
- Questionnaires,
- Interviews,
- Heuristic evaluations and expert reviews.

While the results of evaluations based on most of these methods can be aggregated and presented as individual numerical measures/metrics, the underlying data is typically richer (including not only separate numerical measurements/samples, but also sets, univariate and multivariate series, and qualitative data). Consequently, it should be stated that one-to-one mapping from individual test runs (used for technical performance evaluation) to the tasks and measurements required for socio-technical decision quality evaluation will not be always possible. Furthermore, analyses of the data gathered as part of such evaluations should not be reduced solely to aggregation and summarization as individual metrics, as further quantitative and qualitative analysis methods should be employed to provide insights into the current state of human-AI collaboration and potential improvements for the respective solutions (Lundberg, et al., 2021).

Multiple evaluation metrics will be measured with questionnaires comprising Likert-scale items. While the responses for such items constitute ordinal data that may raise certain methodological concerns with respect to aggregation (e.g., averaging), the prior work provides evidence that results of such analyses are still statistically robust ((Norman, 2010); (Sullivan, et al., 2013)). In case of designing a novel questionnaire, its internal consistency reliability across several dimensions (including the response process validity concerns related to the respondents' understanding and interpretation of survey items, e.g., due to language barriers or use of technical jargon) should be initially assessed with

a pilot study (Rickards, et al., 2012). Additionally, we should acknowledge the potential validity risks with subjective measurements for XAI ((Buçinca, et al., 2020); (Bansal, et al., 2021)) that should be taken into account when planning the studies.

5.2.2 EVALUATION METRICS

The evaluation objectives defined for Task 4.3 in the AI4REALNET project are are based on the project proposal as well as the ongoing project work, such as KPIs already defined for use cases in deliverable D1.1 (AI4REALNET framework and use cases):

- Social-technical decision quality,
- AI acceptability, trust, and trustworthiness,
- Human-user experience,
- AI and human learning curves,
- Task allocation balance,
- Long-term consequences of AI-assistants.

Most of evaluation metrics defined for each of these clusters corresponds to a **special evaluation setup** (see §7 - Execution and reporting) and will rely on objective measurements (such as participant task performance and psychophysiological measurements) and subjective data (such as questionnaires and interviews). As the result of analysis and discussion of KPIs outlined in use cases of D1.1 among project partners, redundancies and considerable overlaps have been identified. Clustering of evaluation metrics thus serves the purpose of prioritizing important recurrent concepts and operationalizing them as part of Task 4.3 evaluation activities and objectives (while further KPIs outlined in D1.1 may be considered part of future work within and beyond of the scope of the project).

In the rest of this section, each evaluation objective and respective cluster are discussed, with the priority given to domain- and use-case-independent evaluation metrics. One important comment to be made here is that while the core measurements for such evaluation metrics, e.g., the number of human interventions, can be indeed considered **domain- and use-case-independent**, the specifics of measurement/calculation and especially the interpretation of the results may be dependent on the context (e.g., the low number of interventions may not represent the significance of required revisions for the AI-proposed solution).

5.2.2.1 SOCIAL-TECHNICAL DECISION QUALITY

This objective puts the stress on *decision quality* within the context of human operator interaction with the AI assistant (cf. “The efficiency of combined human-AI performance” and “Quality of AI-based solutions perceived by human operators” in Annex 4 of D1.1). The evaluation protocols in this cluster represent the number of human operator interventions and scale of necessary revisions for generated decisions.

Evaluation metric	Description	Related Protocols and Concepts in D1.1
Human intervention frequency **KPI-HS-003** (annex 1)	The proportion of instances in which a human operator intervenes in an automated decision-making process. Measured based on number of decision overrides/adjustments by human operators.	"Acceptance score", "Efficiency score", "Prompt demand rate"
Significance of human revisions **KPI-SS-030** (annex 1)	Subjective assessment of necessary revisions for the AI-generated solutions by the human operator. Self-reported by the operator with a Likert-scale question.	"Significance of human revisions"
Perceived decision novelty **KPI-PS-089** (annex 1)	Subjective assessment of nontriviality for the AI-generated solutions by the human operator. Self-reported by the operator with a Likert-scale question.	n.a.

TABLE 6 – SOCIAL-TECHNICAL DECISION QUALITY METRICS

5.2.2.2 AI ACCEPTABILITY, TRUST, AND TRUSTWORTHINESS

This evaluation objective focuses on human operator-oriented indicators of i) trust and acceptability for the AI assistant in general, and ii) agreement with and trustworthiness of specific AI-generated decisions. The notions of "trust", "trustworthiness", "acceptability", and "acceptance" have been mentioned in the project proposal as well as D1.1, however, it should be mentioned, for instance, that Section 3.3 of D1.1 ("Epistemological and Philosophical Foundations of Trustworthy AI") does not provide a universal definition of trustworthiness and states that "… there is no agreement on the determinants of trustworthiness in AI, namely, on what makes an AI system trustworthy". While the precise content and boundaries of these concepts leave room for scientific discussions and future contributions on their own, for purpose of defining evaluation protocols in the context of Task 4.3 these concepts are operationalized as follows:

- Trust is the belief in the reliability, truth, ability, or strength of someone or something. In the context of trust in automation/AI, trust is typically associated with the subjective *attitude* of human users/stakeholders and assessed with questionnaires or interviews (Madsen, et al., 2000), (Spain, et al., 2008).
- Trustworthiness is associated with perceived system reliability and trust for solutions proposed by the system, affecting the user's willingness to act depending on their level of confidence in the system (Madsen, et al., 2000).
- Acceptability is an *a priori* attitude for perceived use that a user has before using the tool, while acceptance is an *a posteriori* pragmatic evaluation related to the actual tool use (Alexandre, et al., 2018).

As part of this objective, the following concerns are addressed:

- Trust and acceptability for the AI assistant in general (see “Acceptance” and “Trust towards the AI tool” evaluation protocols),
- Agreement with AI decisions (see “Agreement score” evaluation protocol),
- Trust for individual AI decisions (see “Trust in AI solutions score” evaluation protocol),
- Perceived decision explainability (see “Comprehensibility” evaluation protocol).

Evaluation metric	Description	Related Protocols and Concepts in D1.1
Acceptance **KPI-AS-002** (annex 1)	Acceptance of the system by a human user assessed with a questionnaire (e.g., based on the Technology Acceptance Model)	“Acceptance”
Trust towards the AI tool **KPI-TS-039** (annex 1)	Self-reported trust (attitude) towards the AI assistant in general assessed with a questionnaire (e.g., the Scale for XAI)	“Trust towards the AI tool”
Agreement score **KPI-AS-005** (annex 1)	Self-reported agreement with individual AI-generated solutions on a scale of 0–100	“Agreement score”
Trust in AI solutions score **KPI-TS-038** (annex 1)	Self-reported trust (attitude) for individual AI-generated solutions on a Likert scale	“Trust in AI solutions score”
Comprehensibility **KPI-CS-013** (annex 1)	Self-reported human operators’ ability to understand and thus make use of the AI-generated decision	“Comprehensibility”

TABLE 7 – AI ACCEPTABILITY, TRUST, AND TRUSTWORTHINESS METRICS

5.2.2.3 HUMAN-USER EXPERIENCE

This evaluation objective focuses on human operator-oriented indicators of user experience, including concerns such as workload and stress; alignment with and support for the operators' cognitive processes; and user motivation and satisfaction. The respective evaluation metrics are measured with quantitative psychophysiological indicators and qualitative assessments.

As part of this objective, the following concerns are addressed:

- Workload and stress (see “Workload”, “Assistant disturbance”, “Cognitive performance and stress” evaluation protocols),

- Alignment and support for the operators' cognitive processes (see “Ability to anticipate”, “Situation awareness”, “Human response time” evaluation protocols),
- User motivation and satisfaction (see “Human motivation”, “Decision support satisfaction” evaluation protocols).

Evaluation metric	Description	Related Protocols and Concepts in D1.1
Workload **KPI-WS-040** (annex 1)	Self-reported human operators’ workload assessed with a questionnaire (e.g., NASA-TLX)	“Workload”, “Workload perception”, “Cognitive load”, “Human Information Processing”
Assistant disturbance **KPI-AS-009** (annex 1)	Self-reported human operators’ assessment of disturbance caused by assistant notifications on a scale [0, 5]	“Assistant disturbance”
Cognitive performance and stress **KPI-CS-049** (annex 1)	Summary of the human operator’s performance status and stress level based on biomarkers	“Cognitive load”, “Human Information Processing”
Ability to anticipate **KPI-AS-001** (annex 1)	Support for human operator's anticipatory sensemaking process assessed with a questionnaire (e.g., Rigor-Metric for Sensemaking)	“Ability to anticipate”
Situation awareness **KPI-SS-031** (annex 1)	Self-reported human operators’ situation awareness measured with a questionnaire (e.g., SAGAT)	“Situation awareness”
Human response time **KPI-HS-023** (annex 1)	Time needed to react to AI advisory/information, measured from the user input	“Human response time”
Human motivation **KPI-HS-022** (annex 1)	Self-reported human operators’ intrinsic work motivation measured with a questionnaire (e.g., based on Job Diagnostic Survey)	“Human motivation”
Decision support satisfaction **KPI-DS-015** (annex 1)	Self-reported human operators’ satisfaction with the system’s support for their decision-making process on a Likert scale	“Decision support for the human operator”, “Decision support satisfaction”

TABLE 8 – HUMAN-USER EXPERIENCE METRICS

5.2.2.4 AI-HUMAN LEARNING CURVES

This evaluation objective focuses on several aspects of the AI and human operator learning, measured with questionnaires.

Evaluation metric	Description	Related Protocols and Concepts in D1.1
AI co-learning capability **KPI-AS-006** (annex 1)	Self-reported human operators' assessment of the AI ability to adapt to the operators' preferences on a Likert scale	"AI co-learning capability"
Human learning **KPI-HS-021** (annex 1)	Self-reported human operators' perceived learning opportunities when working with AI assistant measured with a questionnaire (e.g., based on the task-based workplace learning scale)	"Human learning"

TABLE 9 – AI-HUMAN LEARNING CURVES METRICS

5.2.2.5 AI-HUMAN TASK ALLOCATION BALANCE

This evaluation objective focuses on the optimal balance between AI and human, requirements in terms of new task allocation, measured with human operator-oriented questionnaires.

Evaluation metric	Description	Related Protocols and Concepts in D1.1
Human control/autonomy over the process **KPI-HS-018** (annex 1)	Self-reported human operators' perceived autonomy over the process when working with AI assistant measured with a questionnaire (e.g., based on the Work Design Questionnaire)	"Human control/autonomy over the process"
Impact on workload **KPI-IS-041** (annex 1)	Self-reported human operators' perception of the system impact on their workload (increased/decreased) on a Likert scale	"Workload", "Workload perception"

TABLE 10 – AI-HUMAN TASK ALLOCATION BALANCE METRICS

5.2.2.6 LONG-TERM CONSEQUENCES OF AI ASSISTANTS

This evaluation objective focuses on perceived and predicted long-term consequences of AI assistant adoption. Due to the potential variation in availability of user study participants and stakeholders, the emphasis here is made on evaluation metrics that can be assessed within the scope of individual studies rather than necessarily over a series of experiments; the metrics focusing on reflections over the ongoing/past deployments and pilot studies would thus benefit from being assessed closer to the final rather than initial stages of the project.

Evaluation metric	Description	Related Protocols and Concepts in D1.1
Reflection on operator trust **KPI-RS-091** (annex 1)	Self-reported human operators' perception of the changes in their trust for the AI assistant over time (increased/decreased) on a Likert scale	"Transparency", "Human Agency and Oversight", "Credibility and Intimacy"
Reflection on operator agency **KPI-RS-092** (annex 1)	Self-reported human operators' perception of the changes in their agency working with the AI assistant over time (increased/decreased) on a Likert scale	"Transparency", "Decision support for the human operator", "Human Agency and Oversight"
Reflection on operator de-skilling **KPI-RS-093** (annex 1)	Self-reported human operators' perception of the changes in their own skills working with the AI assistant over time (increased/decreased) on a Likert scale	"Mitigate de-skilling in the human operators"
Reflection on over-reliance **KPI-RS-094** (annex 1)	Self-reported human operators' perception of over-relying on the AI assistant on a Likert scale	"Mitigate addictive behavior from humans"
Reflection on additional training **KPI-RS-095** (annex 1)	Self-reported human operators' perception of the additional training necessary to adopt the AI assistant on a Likert scale	"Additional training about AI for human operators", "Societal and Environmental Well-being"

Evaluation metric	Description	Related Protocols and Concepts in D1.1
Reflection on biases **KPI-RS-096** (annex 1)	Self-reported human operators' perception of biased decisions potentially produced by the AI assistant with respect to gender/ethnicity/age or commercial interests on a Likert scale	"Diversity, Non-discrimination, and Fairness"
Predicted long-term adoption **KPI-PS-097** (annex 1)	Self-reported predicted adoption of the AI assistant by users, stakeholders, or experts on a Likert scale	"Human Agency and Oversight", "Societal and Environmental Well-being"

TABLE 11 – LONG-TERM CONSEQUENCES OF AI ASSISTANTS METRICS

6. ECONOMIC AND REGULATORY ASSESSMENT

This section details contribution of Task 4.4 to the evaluation.

The Task 4.4 is planned to start during M25, i.e., 6 months after the finalization of this report. The following subsections thus rely on initial discussions with the partners involved in the respective task and provide an outline of the assessment to be performed during until M42, which will be subject to active research work conducted by the academic partners involved in Task 4.4.

6.1 ECONOMIC BENEFITS ASSESSMENT

Economic assessment will refer to the approach of benefit evaluation (Miragliotta, 2024). This work is related to the economic evaluation related to the adoption of new technology and can be performed through three approaches:

1. Conditioned evaluation (e.g. ROI is greater than a given threshold),
2. Convincing evaluation (e.g. Perceivable impact on a set of performance),
3. Emotional evaluation.

To have a more rational decision and evaluation, the **conditioned evaluation** has been selected.

6.1.1 ACTIVITY BASED COSTING METHODOLOGY

ABC methodology is a method used to analyze and model the operational processes of a system by linking the activities performed with the necessary resources and associated costs (see (Mahal, et al., 2015), (Quesado, et al., 2021), (Almeida, et al., 2017)). It will be applied to assess the impact of AI solutions in critical contexts of the three selected domains.

The main steps of the methodology to be applied are the following:

5. **Analyze processes and activities**: Understand the key processes of each application domain and decompose them into individual activities.
6. **Link activities to resources**: Gather key execution information for each activity (e.g., allocated resources, timelines, interdependencies with other activities) and estimate the costs associated with resources and activities.
7. **Quantify economic impacts**: Evaluate the efficiency of activities and the impact of introducing new AI solutions and calculate economic indicators such as the total cost of the solution, return on investment (ROI), and the added value of changes (qualitative and quantitative).
8. **Simulate scenarios**: Depending on the data available, sensitivity analysis may be implemented to evaluate the economic outcomes of technology adoption under different scenarios (Worst, Best, Base scenarios).

In the context of the project, the application of this methodology is realized through:

- Constant interaction with the Work Packages (WPs) responsible for developing AI solutions to gather details on model performance and related costs,

- Collaboration with domain partners to collect information for process modelling and data on the costs of introducing AI systems into processes (e.g., staff training, switching costs, change management costs),
- Literature review to identify benefit estimates in comparable contexts, providing benchmarks to compare project results and/or valuable inputs to enhance the estimation model itself.

6.1.2 EVALUATION PROCESS

To conduct an Economic Benefits Assessment for the AI4REALNET project, a structured approach that combines quantitative and qualitative methodologies is required to analyze the economic benefits generated by implementing AI solutions, starting from KPI selection to arrive to quantitative improvement (number based).

First step of the assessment is to **define and analyze the scope of analysis** by studying the operational processes involved and the application contexts (e.g., human decision-making support).

Then, the ABC methodology is applied to compare costs in the **AS-IS scenario** (currently, without AI solution in operation) and **TO-BE scenario** (with AI solution in operation). Estimating costs in AS-IS scenario helps validate potential cost reductions, while estimating costs for the TO-BE scenario provides further input to assess the investment's viability.

Finally, **economic benefits are identified and quantified**, distinguishing tangible benefits (e.g., reduced operational costs, increased efficiency, operational risk savings) from intangible benefits (e.g., improved customer satisfaction, enhanced environmental sustainability performance). This will rely on evaluation outcomes: the impact of the AI systems on selected metrics will be transformed in quantified benefit and evaluated from an economic perspective. Following benefits will be considered:

- Efficiency gains or cost savings derived from automation and optimization,
- Implementation costs and staff training expenses,
- Direct economic benefits such as improved service quality or reduced operational risks,
- Indirect economic benefits such as deployment at larger scales (volume effect), or in other domains related to operation of infrastructures or grids (e.g. public transportation, taxi network, boat and public navigation, port, hydraulic water system, etc.).

Data collection will therefore include:

- **Base zero data**: selected evaluation metrics (e.g. technical performance metrics) will give a quantitative picture of the improved parameters.
- **Primary data**: Through interviews and workshops with project partners and stakeholders, as well as feedback from human operators and managers.
- **Secondary data**: Through benchmarks of similar projects, academic literature, and statistics related to the costs of currently implemented technologies.

To ensure clarity and simplicity, two parameters, Return on Investment (ROI) and cash flow, as suggested by the project guidelines, have been used. ROI prioritizes and highlights the convenience

for stakeholders, while cash flow provides a realistic view of expected financial outcomes before starting any project.

This general approach will then be adapted to each specific application domain and may be modified according to the context.

6.2 REGULATORY ASSESSMENT

The **regulatory assessment** will be mainly based on current applicable EU regulations and ENISA[11] recommendations, and will consider:

- Issues related to ethics, data protection, dataset, algorithm bias, infrastructure vulnerabilities, etc.,
- Impacts on the society, workers, individuals, minorities, etc., focusing on any qualms or worries due to the adoption of AI, and relying on stakeholders' consultation,
- Potential evolution of employment law and workers' rights with AI,
- Other existing ethical, legal, social and impact research methodologies, such as ELSA (Hilten, et al., 2025) or Responsible Research and Innovation[12] (RRI) methodology,
- Relevant work from sister projects (same EU funding programme).

The assessment will be risk based, i.e. impact, consequences and likelihood will be estimated based on experts' opinions (legal and AI experts), considering the innovation of the project, and providing practical experiences in AI model development.

Following the regulatory assessment, a **mitigation plan** will be proposed. This plan will include:

- what is already implemented,
- what is planned to be implemented and the deadline,
- what is not implemented nor planned,
- mitigation actions.

The mitigation plan will be discussed with all relevant stakeholders, from people involved in design of algorithms and interfaces, to the operational experts considering the use of AI systems.

[11] EU Agency for Cybersecurity

[12] See Framework Programmes for Research and Technological Development, funding EU programmes (Horizon)

7. EXECUTION AND REPORTING

This section defines how evaluations are technically setup, run and reported. Furthermore, it defines the project responsibilities during execution. It does not apply to economic and regulatory assessment (see §6).

7.1 CONTEXT

The AI4REALNET project carries out two validation campaigns in sequential order with the aim of validating the outcomes of the project according to a predefined set of evaluation objectives (see Section 2). During both campaigns, metrics will be evaluated for all the three domains (power grid, railway, air traffic management) based on domain specific scenarios (see Sections §3 to §5). The evaluation metrics address a wide range of aspects of the conceptual AI4REALNET framework, from the technical performance of AI models to the robustness against adversarial attacks to the human-AI-interaction experience. Thereby, the digital environment and the human-machine-interface (HMI) play a central role in the calculation of the evaluation metrics. This section describes how the validation campaigns will be executed, both from a procedural and from a technical viewpoint.

7.2 PROCEDURE

The validation campaigns consist of the calculation of the individual evaluation metrics based on their respective evaluation protocols (see sections §3 to §5), carried out following a standardized procedure outlined in this subsection.

Based on the type of result data generation process, an evaluation protocol falls into one of three **Evaluation Protocol Categories:**

i. **Fully automated evaluation:** result data generation using the digital environments, (e.g. evaluation of RL agents' technical performance),
ii. **Semi-automated evaluation:** result data generation from human-computer-interactions using the digital environments and a common human-machine interface, e.g. evaluation of the user experience of handling a scenario by a human operator assisted by an AI,
iii. **Special evaluation setup:** automated, semi-automated or manual evaluation result data generation**,** e.g. evaluation of explanations generated by an AI-model or of mockups.

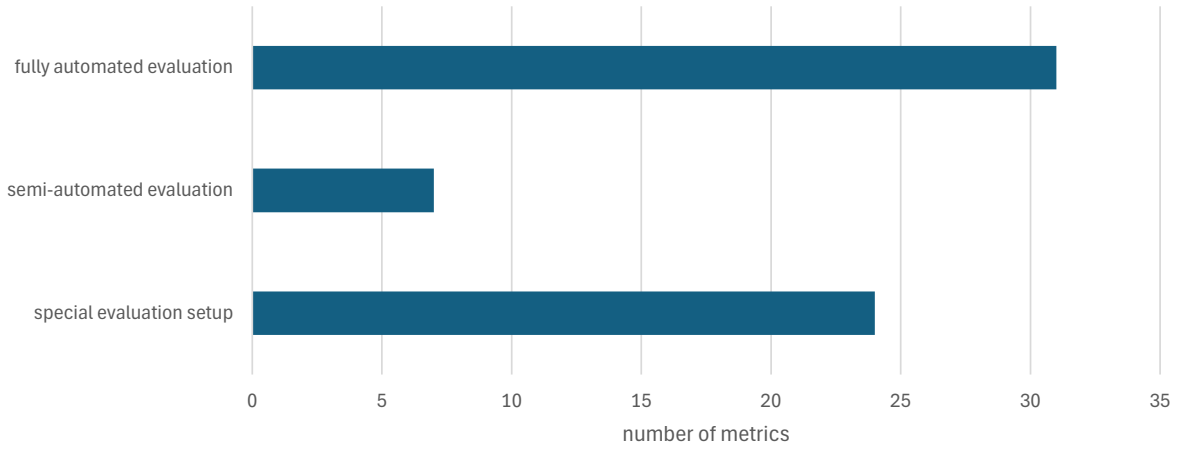


FIGURE 24 – EVALUATION PROTOCOL CATEGORIES

While most of the evaluation protocols fall under category **[i]** and **[iii]** (see Figure 24), some tests require specific equipment or might have restrictions due to data protection regulations (see Data Management Plan). In these cases, they fall under category **[ii]** and the evaluation protocols will be executed in specific testing setups. The identification of a specific setup is based on the description of the evaluation protocols and will be refined, if needed, during the preparation phase as well as after the assessment of the first validation campaign. The final setup will then be used during the second validation campaign.

According to the validation framework depicted in Figure 5, each evaluation objective (see sections 3.2.1, 3.2.2, 4.3.3 and 5.2.2) yields multiple evaluation protocols that are grouped in a **benchmark**. For each evaluation protocol, different tests are run, producing evaluation metrics (KPIs) that are combined in a **single normalized score for the benchmark**. These principles are further detailed hereafter.

For each evaluation protocol, one or multiple tests will be specified during the preparation for the validation campaigns. All tests consist of the following:

- unique identifier,
- evaluation protocol category (**[i]**, **[ii]** or **[iii]**).

For the evaluation protocol categories **[i]** and **[ii]**, tests additionally have the following properties:

- one or multiple scenarios in the form of a script that determines the trajectory of the simulation and will be run using the domain specific digital environment and, in case a human interacts with the scenario directly, in connection with an HMI,
- metrics (KPIs) to be recorded during the test run,
- a flag if manual scoring is required,
- a scoring function that combines one or multiple metrics for each scenario into a **single normalized evaluation score** (KPI) for the test,
- an AI model that determines the actions taken by the agent(s) in the digital environment,

- additional data needed for the scenario run like historical data, predefined events affecting the simulation, etc.

For organizing the evaluation, identifying weak points in reaching the evaluation objectives and determining the success of the final validation campaign, the tests are grouped by evaluation objective of the corresponding evaluation protocol, forming a set of **benchmarks**. A benchmark consists of the following properties:

- tests of the evaluation protocols associated to the objective of the benchmark,
- aggregation function that defines how the scores of the tests are weighted and combined into a **single normalized score for the benchmark.**

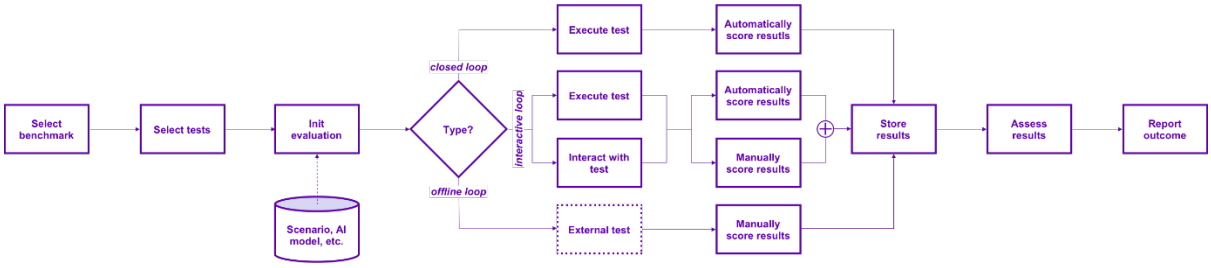


FIGURE 25 – PIPELINE FOR EVALUATION METRICS

A **common evaluation pipeline** will be used to execute one or multiple tests to calculate metrics (see Figure 25 – Pipeline for evaluation):

1. the benchmark is selected,
2. the specific tests that should be executed are selected,
3. the set of tests combined with further data (scenarios, AI model, etc.) are submitted to initialize an evaluation,
4. for each test, one of the following steps are carried out depending on the evaluation protocol category **[i]**, **[ii]** or **[iii]** described above:
    i. for **fully automated evaluation** protocols, the test is executed and scored fully automatically,
    ii. for **semi-automated evaluation** protocols, the test is executed, and an HMI is connected to the test run so that user can interact with the test using the HMI. The test can be scored automatically and/or manually, and the scores are combined using the scoring function of the test,
    iii. for **special evaluation setup** protocols, if a special setup for the test is required, the test is run offline, i.e. will be carried out using the specific setup. Even if the test is run and scoring done fully automatically in this setup, the score will be recorded manually within the pipeline,
5. the results, i.e. test scores as well as additional data generated during the tests, is stored,
6. the results are assessed, i.e. it is checked whether the score reached the defined thresholds.
7. the outcome (results and assessment) is reported.

For the evaluation protocol categories **[i]** (fully automated evaluation) and **[ii]** (semi-automated evaluation), the test execution follows the same logic:

1. a new digital environment is instantiated according to the scenario,
2. the simulation is started and the AI agent(s) and, depending on the test, human users interact with the simulation and affect the state of the environment through their actions,
3. the simulation reaches the end of the scenario run (either by a termination criterion or because the scenario has a fixed sequence of simulation steps) and the trajectory of the run (sequence of environment state, rewards, actions and other model output at each timestep) is exported,
4. the run is scored using the trajectory data based on the scoring function provided by the test.

The tests involving human interactions in **special evaluation setups** are more diverse and cannot be streamlined in the same way. Human-centered evaluation/validation concerns will affect the design of the proposed architecture and require certain flexibility due to the breadth and depth of the respective problems, evaluation methods, and measurements/metrics. The spectrum of methodologies adopted from the behavioral and social sciences (MacGrath, 1995) to human-computer interaction and visualization (Carpendale, 2008) range from field studies, case studies, and sample surveys to laboratory experiments. Evaluations can aim to objectively measure the performance of participants solving specific tasks with a system (typically measuring time and error as part of controlled experiments), elicit subjective feedback and opinions (typically involving interviews and questionnaires), or involve domain experts in qualitative result inspection and discussion (Isenberg, et al., 2013). With respect to evaluation of human-AI interaction and human-centered ML concerns in particular, guidelines have been proposed for assessing AI/ML general properties (such as performance or controllability), explanation properties (such as fidelity or transparency), interface or interaction design, interaction impact, or domain insight (Sperrle, et al., 2021). (Li, et al., 2023) describe a protocol for evaluating human-AI interaction through factorial surveys involving short narratives (vignettes) and corresponding survey questionnaires.(Silva, et al., 2044) contribute a detailed 30-question survey for capturing human-rated simulatability, transparency, and usability of the agent's explanations in the context of human-AI interaction, while (Nauta, et al., 2023) identify 12 conceptual properties (such as correctness, compactness, and coherence) for quantitative evaluation of explanations in XAI. In line with the methods described above, data collection for such studies can involve screen and audio recordings, eye tracking, but also notes and questionnaire results (Sperrle, et al., 2021). The rich data captured from such evaluations can then provide the basis for further analysis of interactions between human operators and AI over time, such as the model supported by the Joint Control Framework (Lundberg, et al., 2021).

7.3 TECHNICAL SETUP

A shared platform, the **Validation Campaign Hub**, to manage all tests and report the results will be set up for all domains, using the **Flatland Benchmarks** (FAB) framework (https://github.com/flatland-association/flatland-benchmarks).

The FAB framework combines concepts from **Codabench** (https://www.codabench.org) and the **Flatland Evaluator** used in the Flatland Challenges (https://www.aicrowd.com/search?q=flatland). In addition to Codabench, FAB will also support the execution of tests involving an HMI as well as manual scoring of tests.

Domain	Benchmark platform	Evaluation backend	Digital environment	HMI
Air traffic management	FAB	BlueSky/FAB	BlueSky	InteractiveAI
Power Grid		Grid2Op/FAB	Grid2Op	
Railway		FAB	Flatland	

TABLE 12 – FRAMEWORKS USED FOR TEST MANAGEMENT, EXECUTION AND EVALUATION

For running the tests and scoring the results, a domain specific digital environment and evaluation backend is used:

- the **BlueSky digital environment** (https://github.com/TUDelft-CNS-ATM/bluesky) in combination with FAB for the air traffic management domain,
- The **Grid2Op digital environment** (https://github.com/Grid2op/grid2op) and evaluator[13] in combination with FAB for the power grid domain,
- the **Flatland digital environment** (https://github.com/flatland-association/flatland-rl) and FAB framework for the railway domain.

For tests requiring an HMI to interact with the test run, the **InteractiveAI framework** (https://github.com/IRT-SystemX/InteractiveAI) will be used which works with all three digital environments. See Table 12 for an overview of all the frameworks.

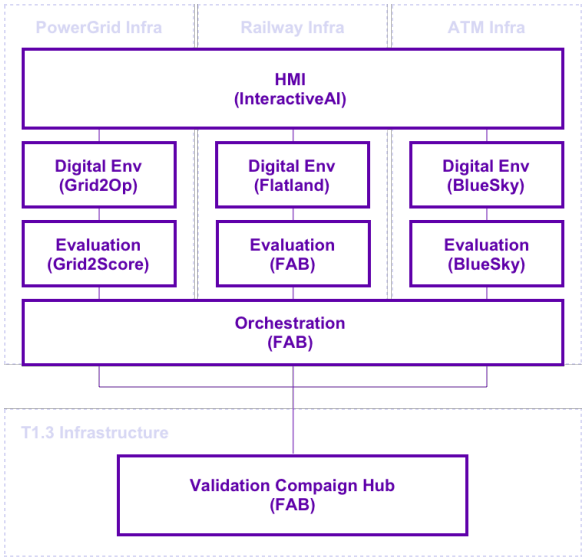


FIGURE 26 – INFRASTRUCTURE AND DEPLOYMENT

While the Validation Campaign Hub will be set up on a common infrastructure shared by all domains (managed by Task 1.3 partners), the evaluation backends including the digital environments will be set up for each domain separately and on dedicated infrastructures (managed by respective domain

[13] This corresponds to the package developed within the project to calculate needed KPIs.

partners) to distribute the computational resource load and to manage access to potentially proprietary or sensitive data.

For tests involving human interactions with the AI system, an HMI will be run on the same infrastructure as the evaluation backend to preserve data protection integrity and to streamline communication between evaluation backend and HMI. See Figure 26 for the deployment setup across infrastructures, and Figure 27 for a domain agnostic view of the evaluation architecture.

Figure 28 shows the information flow between actors (Algorithmic Researcher, Human Factors Researcher, Domain Expert Evaluator and the functional modules involved in the evaluation, Operator) and the functional modules involved in the evaluation (building on the project's Building Block View from the conceptual framework D1.1). The 3 evaluation categories **[i]** fully automated evaluation, **[ii]** semi-automated evaluation and **[iii]** special evaluation setup are reflected by alternative flows and correspond to the three evaluation pipeline loops (closed loop, interactive loop, offline loop) shown in Figure 24.

In the closed and interactive loops, the Test Runner interacts with the simulation (Prediction Module), stepping through simulation time steps for the tests; the output is passed to a Test Evaluator and the Operators.

In the interactive loop, additionally, an Operator interacts with the simulation via Recommendation and Human-Machine Interaction Modules, and the Operator's measures can be taken into account by the Test Runner (reflecting external effects as in Production Systems). The Human Factors Researcher can consider additional feedback from the Operator and a Domain Expert Evaluator.

Finally, in the offline loop, the evaluation results come from an external test setup.

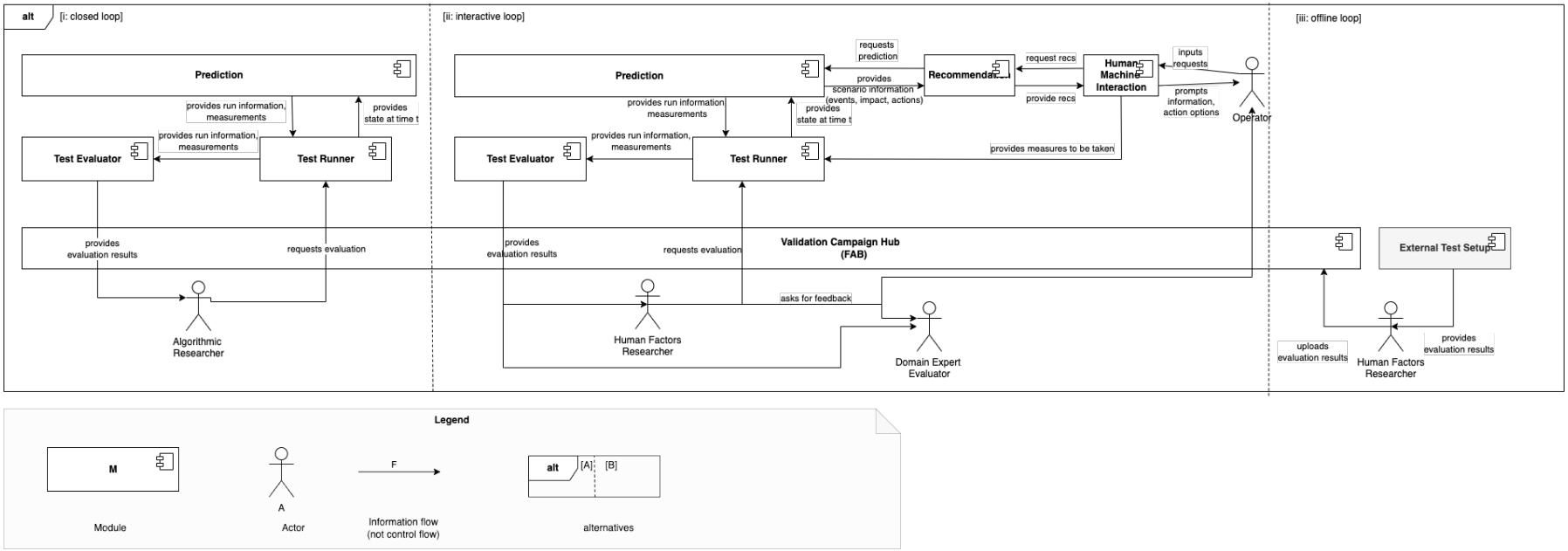


FIGURE 27 – EVALUATION ARCHITECTURE

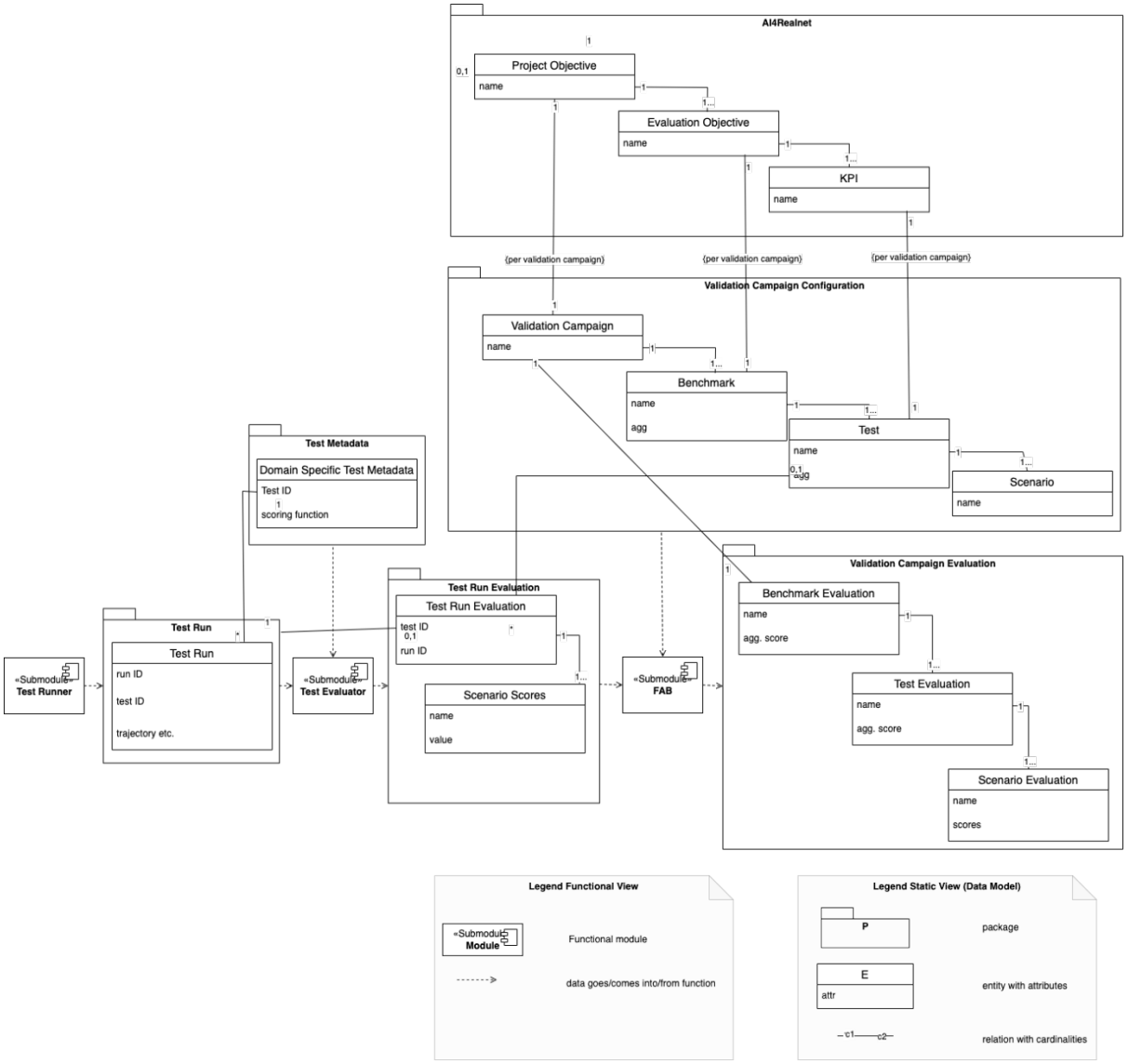


FIGURE 28 – DATA MODEL

Figure 28 gives an overview of the technical setup of evaluations. Referring to the functional modules in Figure 28, it shows how the results from a Test Runner are evaluated by a Test Evaluator and then combined by Benchmark (aggregated score) in the Validation Campaign Hub. Additionally, it shows the underlying data model as described in §7.2 and §7.4. The Test Evaluator uses Test Metadata to score a Test Run. The Validation Campaign Hub aggregates Test Run Evaluations based on its Validation Campaign Configuration.

7.4 ORGANIZATION

7.4.1 COORDINATION

A template for the specification of tests based on the properties outlined in 7.1 will be provided by the validation campaign lead during the preparation phase before the start of the validation campaign. For each KPI, the implemented tests will be tested and verified at the beginning of the campaign by

the partners responsible for the evaluation infrastructure (see Figure 7.2) to ensure smooth execution of the campaign.

Partners responsible for an evaluation metric can initialize evaluation runs through the Validation Campaign Hub and everyone can see the results, scoring and preliminary assessment for each test as soon as a test run is over. This continuous evaluation allows on one hand to further improve evaluation protocols during the campaign as well as to always have an overview of what evaluation was already addressed, ensuring that no evaluation is forgotten.

7.4.2 QUALITY ASSURANCE

All test runs will be logged, and the logs will be available to the person initializing the evaluation. If errors occur, the person initializing the evaluation will be informed. The partners responsible for the Validation Campaign Hub and the domain specific evaluation infrastructures monitor the respective parts continuously during the campaigns.

7.4.3 HANDLING OF RESULTS

All results of the evaluation (scores, trajectories, etc.) will be publicly available on the Validation Campaign Hub. Further, at the end of the second validation campaign, the evaluation results will be made available through the project's Zenodo account, accompanied by scenario data, AI models, and trajectories where allowed, i.e. sensitive and proprietary data will be excluded.

8. CONCLUSION

This deliverable is the result of the first 6 months of WP4 activities. It has allowed to describe:

- 62 evaluation protocols covering 12 evaluation objectives and 3 main project's objectives (see Figure 29 – Number of protocols per evaluation objectives),
- 12 operational scenarios (see Figure 30 – Number of scenarios per domain).

FIGURE 29 – NUMBER OF PROTOCOLS PER EVALUATION OBJECTIVES

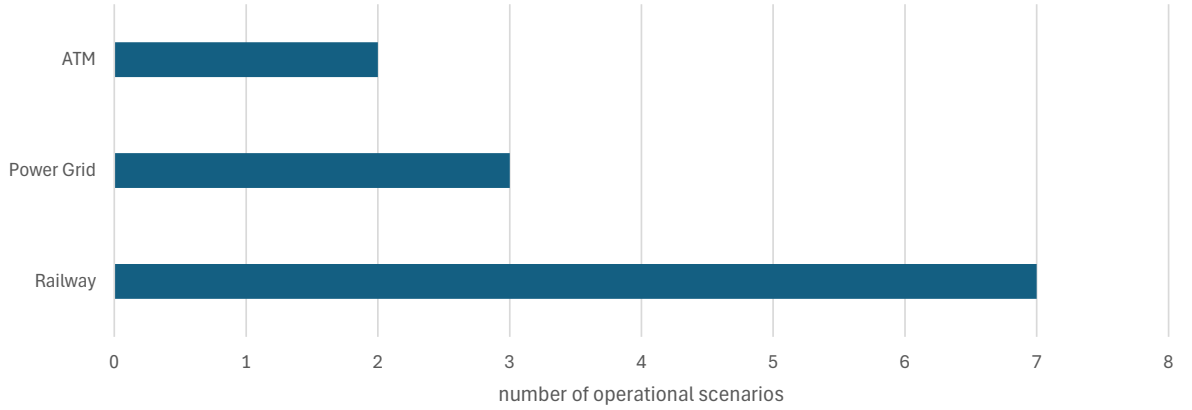


FIGURE 30 – NUMBER OF SCENARIOS PER DOMAIN

The cross-domain opportunities have been used to define a very generic set of evaluation protocols, apart from the technical evaluation of Task 4.1 which is proper to each domain's specificities: all evaluation protocols for safety and robustness (Task 4.2) and social-technical decision quality (Task 4.3) are domain-generic (see Figure 31 – number of evaluation protocols per domain).

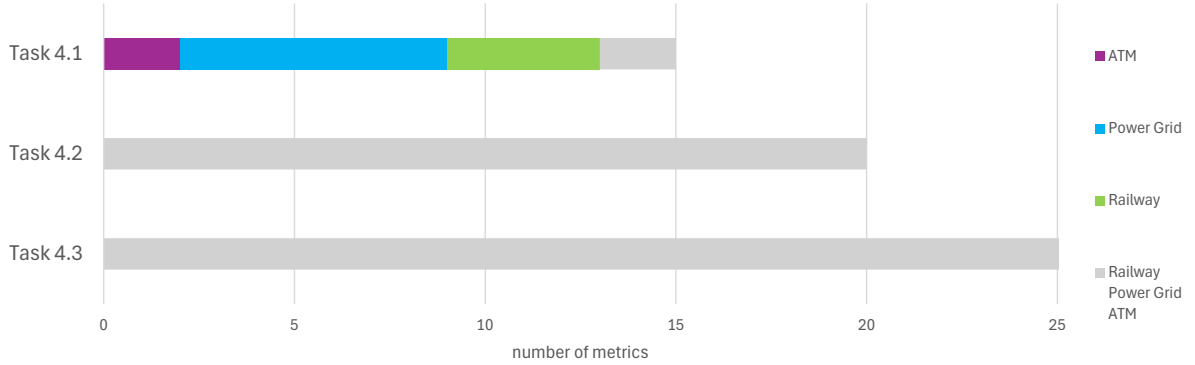


FIGURE 31 – NUMBER OF EVALUATION PROTOCOLS PER DOMAIN

Compared to the other types of evaluation, **social-technical decision quality evaluation** (carried out by Task 4.3) has the highest number of evaluation protocols, which shows the importance of this topic in the planned evaluation work. Especially, the high number of metrics associated to "human-user experience" and "social-technical decision quality" objectives will allow evaluating in detail **what AI can bring on top of current baseline to human operators**: "Perceived decision novelty evaluation metric" can for example show how AI can "augment" the human baseline, by bringing new ways of solving problems. At the same time, evaluation metrics like "Workload", "Assistant disturbance" or "Decision support satisfaction" will show if this isn't done to the detriment of the overall user experience. Socio-technical evaluations of perceived decision novelty (KPI-PS-089), decision support satisfaction (KPI-DS-015), AI co-learning capability (KPI-AS-006), human control/autonomy over the process (KPI-HS-018), and predicted long-term adoption (KPI-RS-097), for instance, will be related to the recent work on complementary improvements from AI augmentation (Bansal, et al., 2021), the perceptions of adaptive autonomous agents in human-AI/robot teams ((Nikolaidis, et al., 2017); (Hoffman, 2019); (Hauptman, et al., 2023)) as well as the impact of AI's adaptability and social role on individual professional efficacy and credit attribution in human–AI collaboration (Du, et al., 2025), thus providing further evidence on the role of AI co-learning for augmenting human operators' potential.

Following finalization of this deliverable, WP4 activities will focus on the preparation of the validation campaign, which include especially the implementation of evaluation protocols (in coordination with WP1 activities), preparation of datasets and technical setup (in coordination with WP2 and WP3 activities).

WP4 activities will also be completed with the start of the last two tasks of the work package: start of Task 4.3 activities in M18 will allow for refining the experimental designs for social-technical evaluation studies. Start of Task 4.4 activities in M25 will allow for setting up and performing the economic and regulatory assessment.

REFERENCES

The following papers, articles and resources are referenced in this document:

Alexandre, Boris, et al. 2018. Acceptance and Acceptability Criteria: A Literature Review. [Online] 2018. https://doi.org/10.1007/s10111-018-0459-1.

Almeida, A. and Cunha, J. 2017. The implementation of an Activity-Based Costing (ABC) system in a manufacturing company. [Online] 2017. https://doi.org/10.1016/j.promfg.2017.09.162.

Alvarez-Melis, David and Jaakkola, Tommi S. 2018. Towards Robust Interpretability with Self-Explaining Neural Networks. [Online] 2018. https://doi.org/10.48550/arXiv.1806.07538.

Amershi, Saleema, et al. 2019. Guidelines for Human-AI Interaction. [Online] 2019. https://doi.org/10.1145/3290605.3300233.

Amodei, Dario, et al. 2016. Concrete Problems in AI Safety. [Online] 2016. https://doi.org/10.48550/arXiv.1606.06565.

Awadi, Afef, Robert, Boris and Langlois, Benoît. 2024. MBSE to Support Engineering of Trustworthy AI-Based Critical Systems. [Online] 2024. https://inrap.hal.science/IRT_SAINT-EXUPERY/hal-04400702v1.

Bansal, Gagan, et al. 2021. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. [Online] 2021. https://doi.org/10.1145/3411764.3445717.

Bedué, Patrick and Fritzsche, Albrecht. 2022. Can We Trust AI? An Empirical Investigation of Trust Requirements and Guide to Successful AI Adoption. [Online] 2022. https://doi.org/10.1108/JEIM-06-2020-0233.

Bessa, Ricardo, Yagoubi, Mouadh and Leyliabadi, Milad. 2024. AI4REALNET Deliverable D1.1. *AI4REALNET framework and use cases.* [Online] 2024. https://ai4realnet.eu/deliverables/.

Brooke, John. 1996. SUS: A 'Quick and Dirty' Usability Scale. [Online] 1996. https://www.taylorfrancis.com/chapters/edit/10.1201/9781498710411-35/sus-quick-dirty-usability-scale-john-brooke.

Buçinca, Zana, et al. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. [Online] 2020. https://doi.org/10.1145/3377325.3377498.

Cahour, Béatrice and Forzy, Jean-François. 2009. Does projection into use improve trust and exploration? An example with a cruise control system. [Online] 2009. https://doi.org/10.1016/j.ssci.2009.03.015.

Carpendale, Sheelagh. 2008. Evaluating Information Visualizations. [Online] 2008. https://doi.org/10.1007/978-3-540-70956-5_2.

Chen, Yize, et al. 2021. Understanding the Safety Requirements for Learning-based Power Systems Operations. [Online] 2021. https://doi.org/10.48550/arXiv.2110.04983.

confiance.ai. 2024. Methodological guideline for Robustness Functional Set. [Online] 2024. https://catalog.confiance.ai/records/km6fw-qsq36.

Corsi, Davide, Marchesini, Enrico and Farinelli, Alessandro. 2021. Formal verification of neural networks for safety-critical tasks in deep reinforcement learning. [Online] 2021. https://proceedings.mlr.press/v161/corsi21a.html.

Du, Tianshu, et al. 2025. Adaptive AI as Collaborator: Examining the Impact of an AI's Adaptability and Social Role on Individual Professional Efficacy and Credit Attribution in Human–AI Collaboration. [Online] 2025. https://doi.org/10.1080/10447318.2025.2462120.

Ebermann, Carolin, Selisky, Matthias and Weibelzahl, Stephan. 2023. Explainable AI: The Effect of Contradictory Decisions and Explanations on Users' Acceptance of AI Systems. [Online] 2023. https://doi.org/10.1080/10447318.2022.2126812.

Endsley, Mica R. 1988. Design and Evaluation for Situation Awareness Enhancement. [Online] 1988. https://doi.org/10.1177/154193128803200221.

—. 1988. Situation awareness global assessment technique (SAGAT). [En ligne] 1988. https://doi.org/10.1109/NAECON.1988.195097.

European Commission. 2023. EU-U.S. Terminology and Taxonomy for Artificial Intelligence. [Online] 2023. https://digital-strategy.ec.europa.eu/en/library/eu-us-terminology-and-taxonomy-artificial-intelligence.

Fitrianie, Siska, et al. 2022. The artificial-social-agent questionnaire: establishing the long and short questionnaire versions. [Online] 2022. https://doi.org/10.1145/3514197.3549612.

Fraser, Doug. 2010. Deskilling: A New Discourse and Some New Evidence. [Online] 2010. https://doi.org/10.1177/103530461002100205.

Frøkjær, Erik, Hertzum, Morten and Hornbæk, Kasper. 2020. Measuring usability: are effectiveness, efficiency, and satisfaction really correlated? [Online] 2020. https://doi.org/10.1145/332040.332455.

Fulton, Nathan and Platzer, André. 2018. Safe Reinforcement Learning via Formal Methods: Toward Safe Control Through Proof and Learning. [Online] 2018. https://doi.org/10.1609/aaai.v32i1.12107.

García, Javier, Majadas, Rubén and Fernández, Fernando. 2020. Learning adversarial attack policies through multi-objective reinforcement learning. [Online] 2020. https://doi.org/10.1016/j.engappai.2020.104021.

Garrido, Josep Soler, et al. 2023. Analysis of the preliminary AI standardisation work plan in support of the AI Act. [Online] 2023. https://data.europa.eu/doi/10.2760/5847.

Goodfellow, Ian J., Shlens, Jonathon and Szegedy, Christian. 2015. Explaining and Harnessing Adversarial Examples. [Online] 2015. https://doi.org/10.48550/arXiv.1412.6572.

Hackman, J. Richard and Oldham, Greg R. 1974. The job diagnostic survey: An instrument for the diagnosis of jobs and the evaluation of job redesign projects. [Online] 1974. https://files.eric.ed.gov/fulltext/ED099580.pdf.

Haider, Tom, et al. 2021. Domain Shifts in Reinforcement Learning: Identifying Disturbances in Environments. [Online] 2021. 10.24406/publica-fhg-412092.

Hauptman, Allyson I., et al. 2023. Adapt and overcome: Perceptions of adaptive autonomous agents for human-AI teaming. [Online] 2023. https://doi.org/10.1016/j.chb.2022.107451.

Hilten, Mireille van, et al. 2025. Ethical, Legal and Social Aspects (ELSA) for AI: An assessment tool for Agri-food. [Online] 2025. https://doi.org/10.1016/j.atech.2024.1007.

Hoffman, Guy. 2019. Evaluating Fluency in Human–Robot Collaboration. [Online] 2019. https://doi.org/10.1109/THMS.2019.2904558.

Hoffman, Robert R., et al. 2023. Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. [Online] 2023. https://doi.org/10.3389/fcomp.2023.1096257.

Hoffman, Robert, et al. 2018. Measuring Trust in the XAI Context. [Online] 2018. http://dx.doi.org/10.31234/osf.io/e3kv9.

Hollnagel, Erik. 2015. Introduction to the Resilience Analysis Grid (RAG). [Online] 2015. https://www.erikhollnagel.com/onewebmedia/RAG%20Outline%20V2.pdf.

Houillon, A., et al. 2013. The effect of novelty on reinforcement learning. [Online] 2013. https://doi.org/10.1016/B978-0-444-62604-2.00021-6.

IEEE. ANSI/ IEEE Std 729-1983. *IEEE standard glossary of software engineering terminology.*

Isenberg, Tobias, et al. 2013. A Systematic Review on the Practice of Evaluating Visualization. [Online] 2013. https://doi.org/10.1109/TVCG.2013.126.

ISO/IEC. 2023. ISO/IEC 23894:2023. *Information technology – Artificial intelligence – Guidance on risk management.* 2023.

—. 2023. ISO/IEC 24029-2:2023. *Artificial intelligence (AI) – Assessment of the robustness of neural networks. Part 2: Methodology for the use of formal methods.* 2023.

—. 2021. ISO/IEC TR 24029-1:2021. *Artificial intelligence (AI) – Assessment of the robustness of neural networks. Part 1: Overview.* 2021.

Kim, Soojong. 2025. Perceptions of discriminatory decisions of artificial intelligence: Unpacking the role of individual characteristics. [Online] 2025. https://doi.org/10.1016/j.ijhcs.2024.103387.

Kosch, Thomas, et al. 2023. A Survey on Measuring Cognitive Workload in Human-Computer Interaction. [Online] 2023. https://doi.org/10.1145/3582272.

Kubicek, Bettina, Paškvan, Matea and Korunka, Christian. 2015. Development and validation of an instrument for assessing job demands arising from accelerated change: The intensification of job demands scale (IDS). [Online] 2015. https://doi.org/10.1080/1359432X.2014.979160.

Lee, Michelle Seng Ah and Singh, Jatinder. 2021. Risk Identification Questionnaire for Detecting Unintended Bias in the Machine Learning Development Lifecycle. [Online] 2021. https://doi.org/10.1145/3461702.3462572.

Lewis, James R., Utesch, Brian S. and Maher, Deborah E. 2013. UMUX-LITE: when there's no time for the SUS. [Online] 2013. https://doi.org/10.1145/2470654.2481287.

Li, Tianyi, et al. 2023. Assessing Human-AI Interaction Early through Factorial Surveys: A Study on the Guidelines for Human-AI Interaction. [Online] 2023. https://doi.org/10.1145/3511605.

Liao, Q. Vera, Gruen, Daniel and Miller, Sarah. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. [Online] 2020. https://doi.org/10.1145/3313831.3376590.

Lu, Jie, et al. 2020. Learning under Concept Drift: A Review. [Online] 2020. https://doi.org/10.48550/arXiv.2004.05785.

Lundberg, Jonas and Johansson, Björn J. E. 2021. A framework for describing interaction between human operators and autonomous, automated, and manual control systems. [Online] 2021. https://doi.org/10.1007/s10111-020-00637-w.

MacGrath, Joseph E. 1995. Methodoly matters: doing research in the behavioral and social sciences. [Online] 1995. https://doi.org/10.1016/B978-0-08-051574-8.50019-4.

Madsen, Maria and Grego, Shirley. 2000. Measuring Human-Computer Trust. [Online] 2000. https://citeseerx.ist.psu.edu/document?doi=b8eda9593fbcb63b7ced1866853d9622737533a2.

Mahal, Ishter and Hossain, Md Akram. 2015. Activity-Based Costing (ABC) – An Effective Tool for Better Management. [Online] 2015. https://www.researchgate.net/publication/309398925_Activity-Based_Costing_ABC_-_An_Effective_Tool_for_Better_Management.

Miragliotta, Giovanni. 2024. *Approach of benefit evaluation.* s.l. : Advanced supply chain lab, Politecnico Milano, 2024.

Morgeson, Frederick P. and Humphrey, Stephen E. 2006. The Work Design Questionnaire (WDQ): Developing and validating a comprehensive measure for assessing job design and the nature of work. [Online] 2006. https://doi.org/10.1037/0021-9010.91.6.1321.

Nasvytis, Linas, et al. 2024. Rethinking Out-of-Distribution Detection for Reinforcement Learning: Advancing Methods for Evaluation and Detection. [Online] 2024. https://doi.org/10.48550/arXiv.2404.07099.

Nauta, Meike, et al. 2023. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. [Online] 2023. https://doi.org/10.1145/3583558.

—. 2023. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. [Online] 2023. https://doi.org/10.1145/3583558.

Nießen, Désirée, et al. 2022. The Internal–External Locus of Control Short Scale–4 (IE-4): A comprehensive validation of the English-language adaptation. [Online] 2022. https://doi.org/10.1371/journal.pone.0271289.

Nikolaidis, Stefanos, Hsu, David and Srinivasa, Siddhartha. 2017. Human-robot mutual adaptation in collaborative tasks: Models and experiments. [Online] 2017. https://doi.org/10.1177/0278364917690593.

Nikolova, Irina, et al. 2014. Work-based learning: Development and validation of a scale measuring the learning potential of the workplace (LPW). [Online] 2014. https://doi.org/10.1016/j.jvb.2013.09.004.

Nilim, Arnab and El Ghaoui, Laurent. 2003. Robustness in Markov Decision Problems with Uncertain Transition Matrices. [Online] 2003. https://papers.nips.cc/paper_files/paper/2003/hash/300891a62162b960cf02ce3827bb363c-Abstract.html.

Norman, Geoff. 2010. Likert scales, levels of measurement and the “laws” of statistics. [Online] 2010. https://doi.org/10.1007/s10459-010-9222-y.

Omnes, Loïc, Marot, Antoine and Donnot, Benjamin. 2021. Adversarial Training for a Continuous Robustness Control Problem in Power Systems. [Online] 2021. https://doi.org/10.48550/arXiv.2012.11390.

Peters, Tobias M. and Visser, Roel W. 2023. The Importance of Distrust in AI. [Online] 2023. https://doi.org/10.1007/978-3-031-44070-0_15.

Pinto, Lerrel, et al. 2017. Robust Adversarial Reinforcement Learning. [Online] 2017. https://proceedings.mlr.press/v70/pinto17a.html.

Quesado, Patricia and Silva, Rui. 2021. Activity-Based Costing (ABC) and Its Implication for Open Innovation. [Online] 2021. https://doi.org/10.3390/joitmc7010041.

Rajeswaran, Aravind, et al. 2017. EPOpt: Learning Robust Neural Network Policies Using Model Ensembles. [Online] 2017. https://doi.org/10.48550/arXiv.1610.01283.

Reisinger, A., et al. 2020. The Concept of Risk in the IPCC Sixth Assessment Report: A Summary of Cross-Working Group Discussions: Guidance for IPCC Authors. [Online] 2020. https://www.ipcc.ch/event/guidance-note-concept-of-risk-in-the-6ar-cross-wg-discussions/.

Rickards, Gretchen, Magee, Charles and Artino Jr, Anthony R. 2012. You Can't Fix by Analysis What You've Spoiled by Design: Developing Survey Instruments and Collecting Validity Evidence. [Online] 2012. https://doi.org/10.4300/JGME-D-12-00239.1.

Ryan, Richard M. and Deci, Edward L. 2000. Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions. [Online] 2000. https://doi.org/10.1006/ceps.1999.1020.

Sadeghi, Fereshteh and Levine, Sergey. 2017. CAD2RL: Real Single-Image Flight without a Single Real Image. [Online] 2017. https://doi.org/10.48550/arXiv.1611.04201.

Scheuer, Dennis. 2020. Entwicklung eines Theoriemodells zur Akzeptanz von Künstlicher Intelligenz. [Online] 2020. https://doi.org/10.1007/978-3-658-29526-4_4.

Schwartz, Reva, et al. 2022. Towards a Standard for Identifying and. [Online] 2022. https://doi.org/10.6028/NIST.SP.1270.

Sedlmeier, Andreas, et al. 2019. Uncertainty-Based Out-of-Distribution Classification in Deep Reinforcement Learning. [Online] 2019. https://doi.org/10.48550/arXiv.2001.00496.

Shneiderman, Ben. 2020. Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. [Online] 2020. https://doi.org/10.1080/10447318.2020.1741118.

Silva, Andrew, et al. 2022. Explainable Artificial Intelligence: Evaluating the Objective and Subjective Impacts of xAI on Human-Agent Interaction. [Online] 2022. https://doi.org/10.1080/10447318.2022.2101698.

Sindermann, Cornelia, et al. 2021. Assessing the Attitude Towards Artificial Intelligence: Introduction of a Short Measure in German, Chinese, and English Language. [Online] 2021. https://doi.org/10.1007/s13218-020-00689-0.

Slattery, Peter, et al. 2024. The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence. [Online] 2024. https://doi.org/10.48550/arXiv.2408.12622.

Spain, Randall D., Bustamante, Ernesto A. and Bliss, James P. 2008. Towards an Empirically Developed Scale for System Trust: Take Two. [Online] 2008. https://doi.org/10.1177/154193120805201907.

Sperrle, F., et al. 2021. A Survey of Human-Centered Evaluations in Human-Centered Machine Learning. [Online] 2021. https://doi.org/10.1111/cgf.14329.

Sullivan, Gail M. and Artino Jr, Anthony R. 2013. Analyzing and Interpreting Data From Likert-Type Scales. [Online] 2013. https://doi.org/10.4300/JGME-5-4-18.

Tapal, Adam, et al. 2017. The Sense of Agency Scale: A Measure of Consciously Perceived Control over One's Mind, Body, and the Immediate Environment. [Online] 2017. https://doi.org/10.3389/fpsyg.2017.01552.

Tobin, Josh, et al. 2017. Domain Randomization for Transferring Deep Neural Networks from Simulation to the Real World. [Online] 2017. https://doi.org/10.48550/arXiv.1703.06907.

Wan, Yu-Yin and Wang, Yi-Shun. 2022. Development and validation of an artificial intelligence anxiety scale: an initial application in predicting motivated learning behavior. [Online] 2022. https://doi.org/10.1080/10494820.2019.1674887.

Yeh, Chih-Kuan, et al. 2019. On the (In)fidelity and Sensitivity for Explanations. [Online] 2019. https://doi.org/10.48550/arXiv.1901.09392.

Zelik, Daniel J., Patterson, Emily S. and Woods, David D. 2018. Measuring Attributes of Rigor in Information Analysis. [Online] 2018. https://doi.org/10.1201/9781315593173-7.

Zheng, Yan, et al. 2021. Vulnerability Assessment of Deep Reinforcement Learning Models for Power System Topology Optimization. [Online] 2021. https://doi.org/10.1109/TSG.2021.3062700.

ANNEX 1 – EVALUATION PROTOCOLS

Page intentionally left blank.

EVALUATION PROTOCOL TEMPLATE

This template has been used to describe all evaluation protocols.

BASIC KPI INFORMATION			
ID	***Application Domain(s)***	***Name of KPI***	
Text	Text	Text	
Version Management			
Version No.	***Date***	***Name of Author(s)***	***Changes***
0.1	DD.MM.YYYY	Text	Text

SCOPE AND OBJECTIVES OF KPI	
Description	The description specifies the KPI and may include specific targets about one of the objectives of the use case and the calculation of these targets. Text
Objective(s)	Here is the link to one of the objectives that are specified in the targets and the KPI. When relevant, KPIs shall be connected to LTEI defined in the project's KPI from AI4REALNET's proposal (not the goal itself), namely: • (LTEI1)S-1, 15%-20% reduction in renewable energy curtailment due to optimal exploration of network flexibility with AI • (LTEI1)S-2, 20%-30% avoided electricity demand shedding • (LTEI1)S-3, 10% increase in punctuality in long-range traffic • 5% increase in punctuality in regional traffic (with realistic disturbances) • (LTEI1)S-4, 3-6% improvement in flight capacity and mile extension Text
Formula(s)	Formula to compute the KPI. When relevant, specify if data must be collected from multiple scenarios, e.g. if the KPI is transversal to different activities: for example, monitor operators while multiple scenarios are run to study cognitive evolution along a set of different scenarios that depend on the system functionalities. Text
Unit of Measurement	Text

CALCULATION METHODOLOGY Describe the procedures used to systematically and reproducibly measure and evaluate the KPIs		
Baseline calculation methodology		
Step #	***Step description***	***Calculation***
1	Detailed description of calculations performed in this step, max. 50 words Text	Actor or tool/module involved in the calculation Text
KPI calculation methodology		
KPI step #	***Step description***	***Calculation***
1	Detailed description of calculations performed in this step, max. 50 words Text	Actor or tool/module involved in the calculation Text

Data collection			
Data ID	***Type***	***Source***	***Frequency***
Text	Description/Name of data type Text	Identify the device, database or source for data collection Text	Indicate how often, when and for how long data is collected Text

LIST OF ALL EVALUATION PROTOCOLS

This table lists all evaluation protocols (ordered by order of appearance in the document) with their relations to project's and evaluation's objectives.

Evaluation protocol	Name	Task	Domain	Project's objective	Evaluation objective
KPI-RF-027	Reduction in delay	Task 4.1	ATM	O2	Effectiveness
KPI-SS-032	System efficiency	Task 4.1	ATM	O2	Effectiveness
KPI-AF-008	Assistant alert accuracy	Task 4.1	Power Grid	O2	Effectiveness
KPI-NF-024	Network utilization	Task 4.1	Power Grid	O2	Effectiveness
KPI-TS-035	Total decision time	Task 4.1	Power Grid	O2	Effectiveness
KPI-DF-016	Delay reduction efficiency	Task 4.1	Railway	O2	Effectiveness
KPI-PF-026	Punctuality	Task 4.1	Railway	O2	Effectiveness
KPI-AF-029	AI Response time	Task 4.1	Railway	O2	Effectiveness
KPI-CF-012	Carbon intensity	Task 4.1	Power Grid	O2	Solution quality
KPI-TF-034	Topological action complexity	Task 4.1	Power Grid	O2	Solution quality
KPI-OF-036	Operation score	Task 4.1	Power Grid	O2	Solution quality
KPI-AS-068	Assistant adaptation to user preferences	Task 4.1	Power Grid	O2	Solution quality
KPI-NF-045	Network Impact Propagation	Task 4.1	Railway	O2	Solution quality
KPI-AF-050	Training time scalability	Task 4.1	ATM Power Grid Railway	O2	Scalability
KPI-AF-051	Test time scalability	Task 4.1	ATM Power Grid Railway	O2	Scalability
KPI-RS-058	Robustness to operator input	Task 4.2	ATM Power Grid Railway	O4	Robustness
KPI-DF-069	Drop-off in reward	Task 4.2	ATM Power Grid Railway	O4	Robustness

Evaluation protocol	Name	Task	Domain	Project's objective	Evaluation objective
KPI-FF-070	Frequency changed output AI agent	Task 4.2	ATM Power Grid Railway	O4	Robustness
KPI-SF-071	Severity of changed output AI	Task 4.2	ATM Power Grid Railway	O4	Robustness
KPI-SF-072	Number of steps in each episode before a critical state is reached	Task 4.2	ATM Power Grid Railway	O4	Robustness
KPI-RF-078	Reward per action	Task 4.2	ATM Power Grid Railway	O4	Robustness
KPI-VF-073	Vulnerability to perturbation	Task 4.2	ATM Power Grid Railway	O4	Robustness
KPI-EF-087	Explainability Faithfulness	Task 4.2	ATM Power Grid Railway	O4	Robustness
KPI-EF-086	Explainability Robustness	Task 4.2	ATM Power Grid Railway	O4	Robustness
KPI-AF-074	Area between the reward curves of the unperturbed and perturbed AI system	Task 4.2	ATM Power Grid Railway	O4	Resilience
KPI-DF-075	Degradation time	Task 4.2	ATM Power Grid Railway	O4	Resilience
KPI-RF-076	Restorative time	Task 4.2	ATM Power Grid Railway	O4	Resilience
KPI-SF-077	Similarity state to unperturbed situation	Task 4.2	ATM Power Grid Railway	O4	Resilience

Evaluation protocol	Name	Task	Domain	Project's objective	Evaluation objective
KPI-DF-057	Domain shift success rate drop	Task 4.2	ATM Power Grid Railway	O4	Reliability
KPI-DF-054	Out-of-domain detection accuracy	Task 4.2	ATM Power Grid Railway	O4	Reliability
KPI-DF-052	Adaptation time and performance drop	Task 4.2	ATM Power Grid Railway	O4	Reliability
KPI-DF-090	Forgetting rate	Task 4.2	ATM Power Grid Railway	O4	Reliability
KPI-DF-056	Robustness to domain parameters	Task 4.2	ATM Power Grid Railway	O4	Reliability
KPI-DF-055	Policy robustness	Task 4.2	ATM Power Grid Railway	O4	Reliability
KPI-DF-053	Generalisation gap	Task 4.2	ATM Power Grid Railway	O4	Reliability
KPI-HS-003	Human intervention frequency	Task 4.3	ATM Power Grid Railway	O3	Social-technical decision quality
KPI-SS-030	Significance of human revisions	Task 4.3	ATM Power Grid Railway	O3	Social-technical decision quality
KPI-PS-089	Perceived decision novelty	Task 4.3	ATM Power Grid Railway	O3	Social-technical decision quality
KPI-AS-002	Acceptance	Task 4.3	ATM Power Grid Railway	O2	AI acceptability, trust, and trustworthiness

Evaluation protocol	Name	Task	Domain	Project's objective	Evaluation objective
KPI-TS-039	Trust towards the AI tool	Task 4.3	ATM Power Grid Railway	O2	AI acceptability, trust, and trustworthiness
KPI-AS-005	Agreement score	Task 4.3	ATM Power Grid Railway	O2	AI acceptability, trust, and trustworthiness
KPI-TS-038	Trust in AI solutions score	Task 4.3	ATM Power Grid Railway	O2	AI acceptability, trust, and trustworthiness
KPI-CS-013	Comprehensibility	Task 4.3	ATM Power Grid Railway	O2	AI acceptability, trust, and trustworthiness
KPI-WS-040	Workload	Task 4.3	ATM Power Grid Railway	O3	Human user experience
KPI-AS-009	Assistant disturbance	Task 4.3	ATM Power Grid Railway	O3	Human user experience
KPI-CS-049	Cognitive performance and stress	Task 4.3	ATM Power Grid Railway	O3	Human user experience
KPI-AS-001	Ability to anticipate	Task 4.3	ATM Power Grid Railway	O3	Human user experience
KPI-SS-031	Situation awareness	Task 4.3	ATM Power Grid Railway	O3	Human user experience
KPI-HS-023	Human response time	Task 4.3	ATM Power Grid Railway	O3	Human user experience
KPI-HS-022	Human motivation	Task 4.3	ATM Power Grid Railway	O3	Human user experience

Evaluation protocol	Name	Task	Domain	Project's objective	Evaluation objective
KPI-DS-015	Decision support satisfaction	Task 4.3	ATM Power Grid Railway	O3	Human user experience
KPI-AS-006	AI co-learning capability	Task 4.3	ATM Power Grid Railway	O3	AI-human learning curve
KPI-HS-021	Human learning	Task 4.3	ATM Power Grid Railway	O3	AI-human learning curve
KPI-HS-018	Human control/autonomy over the process	Task 4.3	ATM Power Grid Railway	O3	AI-human task allocation balance
KPI-IS-041	Impact on workload	Task 4.3	ATM Power Grid Railway	O3	AI-human task allocation balance
KPI-RS-091	Reflection on operator trust	Task 4.3	ATM Power Grid Railway	O3	Long-term consequences of AI assistants
KPI-RS-092	Reflection on operator agency	Task 4.3	ATM Power Grid Railway	O3	Long-term consequences of AI assistants
KPI-RS-093	Reflection on operator de-skilling	Task 4.3	ATM Power Grid Railway	O3	Long-term consequences of AI assistants
KPI-RS-094	Reflection on over-reliance	Task 4.3	ATM Power Grid Railway	O3	Long-term consequences of AI assistants
KPI-RS-095	Reflection on additional training	Task 4.3	ATM Power Grid Railway	O3	Long-term consequences of AI assistants
KPI-RS-096	Reflection on biases	Task 4.3	ATM Power Grid Railway	O3	Long-term consequences of AI assistants

Evaluation protocol	Name	Task	Domain	Project's objective	Evaluation objective
KPI-PS-097	Predicted long-term adoption	Task 4.3	ATM Power Grid Railway	O3	Long-term consequences of AI assistants

TABLE 13 – LIST OF ALL EVALUATION PROTOCOLS

TECHNICAL PERFORMANCE AND SCALABILITY

This annex details all evaluation protocols corresponding to §3 - Technical performance and scalability, and Effectiveness, Solution quality, Scalability objectives.

Page intentionally left blank.

EFFECTIVENESS

BASIC KPI INFORMATION		
ID	**Application Domain(s)**	**Name of KPI**
KPI-RF-027	ATM	Reduction in delay

Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	05.02.2025	J. Soares (NAV)	Creation of the document
0.1	03.03.2025	J. Sоares (NAV)	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	The reduction in delay KPI aims to quantify the time gained overall and for each airplane, with the introduction of AI.
Objective(s)	This KPI aims to quantify the efficiency gains of AI integration by measuring how AI impacts execution time and delays. Specifically, it helps determine whether AI: • Reduces execution time deviations ($diff_{ai} < diff_h$). • Minimizes delays ($delay_{ai} < delay_h$). • Enhances consistency and reliability in operations. By evaluating these metrics, we can assess the AI's effectiveness in improving human decision-making, reducing intervention time, and optimizing operational workflows. This KPI contributes to evaluating Effectiveness of the AI-based assistant, as part of Task 4.1 evaluation objectives, and O2 main project objective. This KPI is linked with project's Long Term Expected Impact (LTEI) (LTEI1)$KPI_{S\text{-}4}$, 3-6% improvement in flight capacity and mile extension.
Formula(s)	**Performance Deviation** measures the percentage deviation of actual time from expected time: $$diff_i = \frac{Ta_i - Te_i}{Te_i} \times 100$$ **Delay Measurement** measures the absolute delay in arrival time: $$delay_j = Taar_j - Tear_j$$ These formulas will be applied to both human-only performance and human-AI collaborative performance, resulting in: • $diff_h$ and $delay_h$ (Human performance) • $diff_{ai}$ and $delay_{ai}$ (Human-AI performance)
Unit of Measurement	Percentage and seconds

CALCULATION METHODOLOGY		
Baseline calculation methodology		
Step #	**Step description**	**Calculation**
1	Use different simulations with representative data.	Digital environment KPI module
2	Identify variables: Expected controlled time and expected arrival time and create the corresponding array.	Digital environment KPI module

3	Take the baseline delay of all human operators for each simulation.	Digital environment KPI module

KPI calculation methodology		
KPI step #	***Step description***	***Calculation***
1	Calculate the delay time of the human-AI team for each simulation.	Digital environment KPI module
2	Create the array with the relative response time outcome for each simulation.	Digital environment KPI module
3	Compute the results using the formula.	Digital environment KPI module
4	Compute the metrics defined in the formula section for each calculated array.	Digital environment KPI module
5	Identify the distribution that best fits each variable.	n.a.
6	Compute median, mean, max, and min for each metric.	n.a.
7	Compare performance: The median, mean, max, and min values for human-AI performance should be lower than for human-only performance, indicating improved efficiency.	Digital environment KPI module

Data collection			
Data ID	***Type***	***Source***	***Frequency***
Te	Expected time	Digital environment	Each aircraft in the supervisor's scope
Ta	Actual time	Digital environment	Each aircraft in the supervisor's scope
$Tear$	Expected arrival time	Digital environment	Each aircraft in the supervisor's scope
$Taar$	Actual arrival time	Digital environment	Each aircraft in the supervisor's scope

BASIC KPI INFORMATION			
ID	**Application Domain(s)**	**Name of KPI**	
KPI-SS-032	ATM	System efficiency	
Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	31.01.2025	J. Soares (NAV)	Creation of the document
1.0	03.03.2025	J. Soares (NAV)	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	System efficiency measures the efficiency of the system in delivering trustworthy solutions requiring less effort and time to deliver an appropriate response by the operator.
Objective(s)	The System efficiency KPI aims to evaluate the effectiveness of the AI solution in real operational conditions, considering not just its raw response time but also the quality and usability of its assistance. This includes how the AI presents its advice, its ease of use, the accuracy of its recommendations, and how well it integrates with existing data and workflows. The evaluation will measure the AI-human collaboration, focusing on: • **Response efficiency**: The time taken for the AI to generate advice and for the human operator to act on it. • **Advice clarity & usability**: How well structured, coherent, and understandable the AI's suggestions are. • **Data integration quality**: How seamlessly the AI incorporates relevant information into its recommendations. • **Human correction factor**: Whether and how often the operator needs to correct or refine the AI's output. • **Decision-making speed**: The overall reduction in response time achieved through AI-assisted operation. By considering these factors, the tests aim to assess how well the AI minimizes human intervention while maximizing efficiency, accuracy, and usability in decision-making. This KPI contributes to evaluating Effectiveness of the AI-based assistant, as part of Task 4.1 evaluation objectives, and O2 main project objective.
Formula(s)	$\boldsymbol{X_i}$ – is the vector that associates each test to the outcome of $Th > Tai$. Meaning that x_i will be, for example, [1, 1, 1, 0, 0, …, x_n], where 1 means that $Th > Tai$, and zero the opposite. $$n = \frac{\sum_{i=1}^{N} x_i}{N} \times 100$$ This indicates that we want to measure how many times the Tai is less than Th in all tests in order to compute the relative efficiency of the tool. The objective is to reach a level of n > 75% and total sum of the times $Th > Tai$.
Unit of Measurement	Percentage (%)

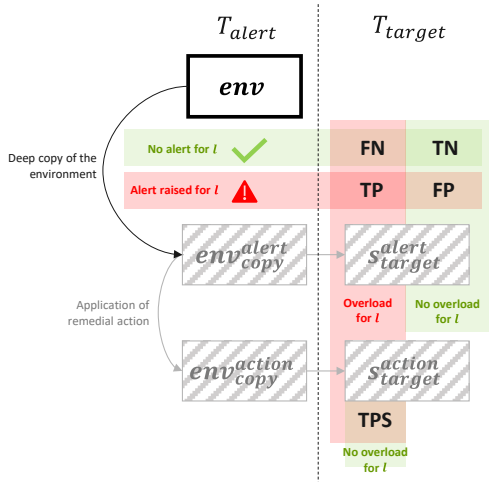
CALCULATION METHODOLOGY		
Baseline calculation methodology		
Step #	**Step description**	**Calculation**

1	Identify all testing scenarios	Simulation and Testing Tools (operations Testing Team)
2	Take the baseline response of all human operators for each testing scenarios.	Digital environment KPI module

KPI calculation methodology		
KPI step #	***Step description***	***Calculation***
1	Calculate the response time of the human-AI team for each testing scenarios.	Digital environment KPI module
2	Create the array with the relative response time outcome for each test.	Digital environment KPI module
3	Apply the formula to get the relative efficiency of the tool.	Digital environment KPI module

Data collection			
Data ID	***Type***	***Source***	***Frequency***
N	Number of tests under different conditions	Digital environment	Once
Th	Time it takes the human to compute a solution	Human Machine Interaction module	Each test
Tai	Time it takes for the AI to compute a solution and the human operator to accept the solution	Human Machine Interaction module	Each test

BASIC KPI INFORMATION			
ID	**Application Domain(s)**	**Name of KPI**	
KPI-AF-008	Power Grid	Assistant alert accuracy	
Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	21.01.2025	J. Viebahn (TenneT)	Creation of the document
0.2	10.02.2025	B. Lemetayer (RTE)	Addition of calculation details
1.0	03.03.2025	B. Lemetayer (RTE)	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	Assistant alert accuracy is based on the number of times the AI assistant agent is right about forecasted issues ahead of time. Even if forecasted issues concern all events that lead to a grid state out of acceptable limits (set by operation policy), use cases of the project focus on managing overloads only: this KPI therefore only **focuses on alerts related to line overloads**. The calculation of KPI relies on simulation of 2 parallel paths (starting from the moment the alert is raised): • Simulation of the "do nothing" path, to assess the truth values • Application of remedial actions to the "do nothing" path, to assess solved cases  To calculate the KPI, all interventions by an agent or operator are fixed to a specific plan since every alert is related to a specific plan (e.g. remedial actions). *Note: line contingencies for which alerts can be raised are the lines that can be attacked in the environment (env.alertable_line_ids in grid2Op), so this should be properly configured beforehand.*
Objective(s)	This KPI contributes to evaluating Effectiveness of the AI-based assistant, as part of Task 4.1 evaluation objectives, and O2 main project objective.

SCOPE AND OBJECTIVES OF KPI	
Formula(s)	The formula to compute the KPI is the confusion matrix (see Calculation Methodology): • **TP, True positive cases**, forecast alerts were raised by the AI assistant, and overloads did occur on the transmission grid) • **FP, False positive cases**, forecast alerts were raised by the AI assistant, but no overload occurred on the transmission grid • **TN, True negative cases**, the AI assistant raised no forecast alert, and no overload occurred on the transmission grid • **FN, False negative cases**, the AI assistant raised no forecast alert, but overloads occurred on the transmission grid Starting from True positive cases, **TPS, the True positive cases solved**, represent the alert effectively solved by the recommendations. The KPI can be computed per episode, across several episodes of one scenario, or even across scenarios.
Unit of Measurement	None (counting)

CALCULATION METHODOLOGY		
Baseline calculation methodology		
Step #	***Step description***	***Calculation***
1	Same as KPI calculation methodology.	Digital environment KPI module
KPI calculation methodology		
KPI step #	***Step description***	***Module(s)***
1	The AI assistant raises an overload alert to the operator at T_{alert}, concerning T_{target} time step, concerning $l_{overload}$ lines: these lines are listed in $\langle l_{overload}^{forecast} \rangle_{T_{target}}^{T_{alert}}$	Digital environment KPI module
2	Starting from $env_{copy}^{T_{alert}}$, a deep copy of the environment made at T_{alert}, the state of environment obtained at T_{target} time step, $s_{T_{target}}^{T_{alert}}$, is calculated.	Digital environment KPI module
3	Starting from env_{copy}^{action}, a deep copy of the environment made at T_{alert}, the remedial actions associated to the alert are applied. Then the state of environment obtained at T_{target} time step, $s_{T_{target}}^{action}$, is calculated.	Digital environment KPI module
4	An overload occurs at T_{target} time step in $s_{T_{target}}^{T_{alert}}$, concerning $l_{overload}$ lines: these lines are listed in $\langle l_{overload} \rangle_{T_{target}}^{T_{alert}}$	Digital environment KPI module
5	Compute $TP_{T_{target}}^{T_{alert}}$, **True positive cases** of T_{target} time step seen from T_{alert} time step, as follows: Number of $l_{overload}$ elements present in $\langle l_{overload}^{forecast} \rangle_{T_{target}}^{T_{alert}}$ and in $\langle l_{overload} \rangle_{T_{target}}^{T_{alert}}$	Digital environment KPI module
6	Compute $TPS_{T_{target}}^{T_{alert}}$, **True positive cases solved**, as number of lines without overload in $s_{T_{target}}^{action}$ that are present in $\langle l_{overload}^{forecast} \rangle_{T_{target}}^{T_{alert}}$ and in $\langle l_{overload} \rangle_{T_{target}}^{T_{alert}}$.	Digital environment KPI module

7	Compute $FP_{T_{target}}^{T_{alert}}$, **False positive cases** of T_{target} time step seen from T_{alert} time step, as follows: Number of $l_{overload}$ elements present in $\langle l_{overload}^{forecast} \rangle_{T_{target}}^{T_{alert}}$ and not in $\langle l_{overload} \rangle_{T_{target}}^{T_{alert}}$	Digital environment KPI module
8	Compute $FN_{T_{target}}^{T_{alert}}$, **False negative cases** of T_{target} time step seen from T_{alert} time step, as follows: Number of $l_{overload}$ elements present in $\langle l_{overload} \rangle_{T_{target}}^{T_{alert}}$ and not in $\langle l_{overload}^{forecast} \rangle_{T_{target}}^{T_{alert}}$	Digital environment KPI module
9	Compute $TN_{T_{target}}^{T_{alert}}$, **True negative cases** of T_{target} time step seen from T_{alert} time step, as follows: $TN_{T_{target}}^{T_{alert}} = n_{alertable} - TP_{T_{target}}^{T_{alert}} - FP_{T_{target}}^{T_{alert}} - FN_{T_{target}}^{T_{alert}}$	Digital environment KPI module
10	Confusion matrix is represented with following vector: $\langle TP, FP, TN, FN, TPS \rangle_{T_{target}}^{T_{alert}}$	Digital environment KPI module

Data collection			
Data ID	***Type***	***Source***	***Frequency***
T_{alert}	Time when overload alert is created	Digital environment	Each observation of the episode
T_{target}	Time step at which the alerted issue is predicted to happen	Digital environment	Each observation of the episode
$\langle l_{overload} \rangle_T^{alert}$	List of lines concerned by the raised alert for a given time step T	Digital environment	Each observation of the episode
$\langle l_{overload} \rangle_T$	List of lines affected by overloads at target time step for a given time step T	Digital environment	Each observation of the episode
env_{copy}^{alert}	Deep copy of the environment performed when overload alert	Digital environment	Each observation of the episode
env_{copy}^{action}	Deep copy of the environment performed when overload alert	Digital environment	Each observation of the episode
$n_{alertable}$	Number of lines for which an alert can be raised	Digital environment	Environment settings

BASIC KPI INFORMATION			
ID	**Application Domain(s)**	**Name of KPI**	
KPI-NF-024	Power Grid	Network utilization	
Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	17.01.2025	B. LEMETAYER (RTE)	Creation of the document
1.0	03.03.2025	B. Lemetayer (RTE)	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	Network utilization KPI is based on the relative line loads of the network, indicating to what extent the network and its components are utilized.
Objective(s)	This KPI contributes to evaluating Effectiveness of the AI-based assistant, as part of Task 4.1 evaluation objectives, and O2 main project objective.
Formula(s)	This KPI yield a vector with 6 values, that are calculated over all scenarios' steps: • the maximum line's load in N state, • the maximum line's load in N-1 state, • the average of the maximum line's load in N state per step, • the average of the maximum line's load in N-1 state per step, • the share of lines where the line's load in N state is greater than 90%, • the share of lines where the line's load in N-1 state is greater than 100%. Line's load is referred to as rho in Grid2Op and is defined as the observed current flow divided by the thermal limit of the line.
Unit of Measurement	Vector of 6 values expressed in percent (decimal number between 0% and 100%)

CALCULATION METHODOLOGY		
Baseline calculation methodology		
Step #	**Step description**	**Module(s)**
1	Apply all KPI calculation steps and calculate the KPI for a baseline agent on the same environment: ideally this should be close to the human behavior	Digital environment KPI module
KPI calculation methodology		
KPI step #	**Step description**	**Module(s)**
1	Get the vector of rho values for all lines for each observation obs in Grid2Op, noted as: $\langle\langle rho^{N}_{line,obs}\rangle_{line}\rangle_{obs}$ Use *obs.rho* vector in Grid2Op.	Digital environment
2	Compute the maximum rho value from vector obtained at step 1: $rho^{N}_{max} = max_{line,obs}\left(\langle\langle rho^{N}_{line,obs}\rangle_{line}\rangle_{obs}\right) \times 100$	Digital environment KPI module

3	Perform a contingency analysis for each observation obs in Grid2Op to get the vector of rho values for all lines in N-1 states of this observation, noted as: $\langle\langle\langle rho^{N-1}_{line,obs,contingency}\rangle_{line}\rangle_{contingency}\rangle_{obs}$ **For the moment, the contingency list to be taken into account is the one with all lines only.** This can be done by multiple means: • Using the built-in functions of the simulator backend, e.g. security analysis of LightSim2Grid: Contingency Analysis (doc in progress) — LightSim2Grid 0.10.0 documentation • Use a *Simulator*, see documentation for Grid2Op *Simulator* class: Simulator — Grid2Op 1.10.4 documentation *Note: To be further updated with function of PyPowsybl backend*	Digital environment KPI module
4	Compute the maximum rho value for each observation obs from step 3: $rho^{N-1}_{max} = max_{line,contingency,obs}\left(\langle\langle\langle rho^{N-1}_{line,obs,contingency}\rangle_{line}\rangle_{contingency}\rangle_{obs}\right) \times 100$	Digital environment KPI module
5	Using vector obtained at step 1, calculate the average of the maximum line's load in N state per step: $rho^{N}_{avg} = avg\left(\langle max_{line}\left(\langle\langle rho^{N}_{line,obs}\rangle_{line}\rangle_{obs}\right)\rangle\right) \times 100$	Digital environment KPI module
6	Using vector obtained at step 3, calculate the average of the maximum line's load in N-1 state per step: $rho^{N-1}_{avg} = avg\left(\langle max_{line,contingency}\left(\langle\langle\langle rho^{N}_{line,obs,contingency}\rangle_{line}\rangle_{contingency}\rangle_{obs}\right)\right) \times 100$	Digital environment KPI module
7	Using vector obtained at step 1, calculate the share of lines where the line's load in N state is greater than 100%: $overload^{N} = \frac{sum_{line,obs}\left(\langle\langle rho^{N}_{line,obs}\rangle_{line}\rangle_{obs} > 1\right)}{n_{line} \times n_{obs}} \times 100$	Digital environment KPI module
8	Using vector obtained at step 3, calculate the share of lines where the line's load in N-1 state is greater than 100%: $overload^{N-1} = \frac{sum_{line,contingency,obs}\left(\langle\langle\langle rho^{N-1}_{line,obs,contingency}\rangle_{line}\rangle_{contingency}\rangle_{obs} > 0.9\right)}{n_{line} \times n_{obs}} \times 100$	Digital environment KPI module
9	Calculate the KPI as follows: $\left\langle \begin{matrix} rho^{N}_{max} \\ rho^{N-1}_{max} \\ rho^{N}_{avg} \\ rho^{N-1}_{avg} \\ overload^{N} \\ overload^{N-1} \end{matrix} \right\rangle$	Digital environment KPI module

Data collection			
Data ID	**Type**	**Source**	**Frequency**
obs	Observation	Grid2Op environment observation	Each episode step
rho	Observation	Grid2Op environment observation property	Each episode step

BASIC KPI INFORMATION		
ID	**Application Domain(s)**	**Name of KPI**
KPI-TS-033	Power Grid	Total decision time

Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	21.01.2025	J. Viebahn (TenneT)	Creation of the document
0.2	05.02.2025	B. Lemetayer (RTE)	Addition of data collection sources
0.3	06.02.2025	J. Viebahn (TenneT)	Addition of alert-to-problem measurement
0.4	06.02.2025	B. Lemetayer (RTE)	Details of calculation steps and data collection sources
1.0	03.03.2025	B. Lemetayer (RTE)	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	It is based on the overall time needed to decide, thus including the respective time taken by the AI assistant and human operator. This KPI can be detailed to specifically distinguish the time needed by the AI assistant to provide a recommendation. An assumption is that a **Human Machine Interaction (HMI) module is available**.
Objective(s)	This KPI addresses the following objectives: *1)* Given an alert, how much time is left until the problem occurs? *The longer the better since it gives more time to make a decision.* *2)* Given an alert, how much time does the AI assistant take to come up with its recommendations to mitigate the issue? *The shorter the better.* *3)* Given the recommendations by the AI assistant, how much time does the human operator take to make a final decision? *The shorter the better since it indicates that the recommendations are clear and convincing for the human operator.* In case there is no interaction possible between the AI assistant and the human operator, this overall split is not possible. Then there is only one overall time needed to decide, starting from the alert and ending with the final decision by the human operator. This KPI contributes to evaluating Effectiveness of the AI-based assistant, as part of Task 4.1 evaluation objectives, and O2 main project objective.
Formula(s)	See KPI calculation methodology
Unit of Measurement	Time (minutes or seconds)

CALCULATION METHODOLOGY		
Baseline calculation methodology		
Step #	**Step description**	**Module(s)**
1	Same as KPI calculation methodology.	Digital environment KPI module
KPI calculation methodology		
KPI step #	**Step description**	**Module(s)**

1	The AI assistant raises an alert to the operator. Measure the time left before the alerted problem is supposed to happen as follows: $$D_1 = T_{target} - T_{alert}^{display}$$ In case there is no HMI module, D_1 is calculated as follows: $$D_1 = T_{target} - T_{alert}$$	Digital environment KPI module
2	The AI assistant raises an alert to the operator. Measure the time it takes for the AI assistant to display the recommendations to mitigate the problem as follows: $$D_2 = T_{recommendation}^{display} - T_{alert}^{display}$$ In case there is no HMI module, D_2 is not calculated	Digital environment KPI module
3	The AI assistant displays the recommendations to the operator. Measure the time it takes for the human operator to make a decision as follows: $$D_3 = T_{decision} - T_{recommendation}^{display}$$ In case there is no HMI module, D_3 cannot be calculated.	Digital environment KPI module
4	Calculate the KPI as follows: $$\langle D_1 \\ D_2 \\ D_3 \rangle$$	Digital environment KPI module

Data collection			
Data ID	***Type***	***Source***	***Frequency***
T_{alert}	Time when alert is created	Digital environment	Each observation of the episode
T_{target}	Time step at which the alerted issue is predicted to happen	Digital environment	Each observation of the episode
$T_{alert}^{display}$	Time when alert is displayed to the operator	Human Machine Interaction module	Each observation of the episode
$T_{recommendation}^{display}$	Time when recommendations are displayed to the operator	Human Machine Interaction module	Each observation of the episode
$T_{decision}$	Time when the operator makes a decision, corresponding to the time when the operator chooses a recommendation	Human Machine Interaction module	Each observation of the episode

BASIC KPI INFORMATION			
ID	**Application Domain(s)**	**Name of KPI**	
KPI-DF-016	Railway Network	Delay Reduction Efficiency	
Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	07.02.2025	Roman Liessner (DB)	Creation of the document
1.0	03.03.2025	Roman Liessner (DB)	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	The Delay Reduction Efficiency KPI quantifies the effectiveness of the AI-driven re-scheduling system in reducing overall train delays. By comparing delays before and after AI intervention, this metric provides insight into the system's capability to optimize train schedules and minimize disruptions.
Objective(s)	This KPI contributes to evaluating Effectiveness of the AI-based assistant, as part of Task 4.1 evaluation objectives, and O2 main project objective: • To assess the impact of AI-based re-scheduling on reducing delays in railway operations. • To ensure that AI interventions lead to measurable improvements in punctuality. • To provide a performance benchmark for AI-driven traffic management solutions in railway networks.
Formula(s)	$DelayReductionEfficiency = \frac{\sum delay_{beforeAI} - \sum delay_{afterAI}}{\sum delay_{beforeAI}} \times 100$
Unit of Measurement	Percentage (%) reduction in total delay time.

CALCULATION METHODOLOGY		
Baseline calculation methodology		
Step #	**Step description**	**Calculation**
1	Collect historical train delay data before AI implementation.	Data Analysts
2	Establish baseline statistics for train delays across different operational scenarios.	Railway Control Center
KPI calculation methodology		
KPI step #	**Step description**	**Calculation**
1	Monitor train schedules and log delay times before and after AI-based re-scheduling	AI System & Control Center
2	Calculate total cumulative delay for both pre-AI and post-AI operation:$\sum delay_{beforeAI}$, $\sum delay_{afterAI}$	Data Analysts
3	Compute the delay reduction efficiency using the formula above.	Data Analysts
4	Analyze trends and refine AI system parameters to improve efficiency.	Railway Operators

Data collection			
Data ID	**Type**	**Source**	**Frequency**
$delay_{beforeAI}$	delay before AI-based re-scheduling	Train schedules	End of episode
$delay_{afterAI}$	delay after AI-based re-scheduling	Train schedules	End of episode

BASIC KPI INFORMATION			
ID	**Application Domain(s)**	**Name of KPI**	
KPI-PF-026	Railway	Punctuality	
Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	20.01.2015	Roman Liessner (DB)	Creation of the document
1.0	03.03.2025	Roman Liessner (DB)	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	Punctuality measures the percentage of trains arriving at their destinations on time **(the train doesn't arrive after planned arrival) and the train didn't depart before planned departure time**. The goal is to maintain a high level of reliability and minimize delays for passengers and freight services.
Objective(s)	This KPI contributes to evaluating Effectiveness of the AI-based assistant, as part of Task 4.1 evaluation objectives, and O2 main project objective: • Improve customer satisfaction by ensuring timely arrivals • **Guarantee maximal planned connection** • Minimize operational disruptions caused by delays • Meet regulatory and stakeholder benchmarks for punctuality This KPI is linked with project's Long Term Expected Impacts (LTEI) (LTEI1)$KPI_{S\text{-}3}$: • 10% increase in punctuality in long-range traffic • 5% increase in punctuality in regional traffic (with realistic disturbances)
Formula(s)	$PunctualityRate = \frac{n_{arrivals}^{onTime}}{n_{arrivals}} \times 100$
Unit of Measurement	Percentage (%)

CALCULATION METHODOLOGY			
Baseline calculation methodology			
Step #	**Step description**		**Module(s)**
1	Determine the total number of arrivals and on-time arrivals during the period **under the condition the train didn't depart too early.**		Timetable Analysis Tool (Operations Data Analyst)
KPI calculation methodology			
KPI step #	**Step description**		**Module(s)**
1	Extract data from train operation logs		Monitoring System (Data Engineer)
2	Validate data integrity, ensuring consistency in defining "on-time" arrivals (e.g., within 6 minutes of scheduled arrival).		Data Cleaning Algorithms (Data Scientist)
3	Calculate punctuality rate using the formula.		Analytical Reporting System (Operations Analyst)
Data collection			
Data ID	**Type**	**Source**	**Frequency**
$n_{arrivals}$	Total number of arrivals	train operation logs	End of episode
$n_{arrivals}^{onTime}$	Number of on time arrivals	train operation logs	End of episode

BASIC KPI INFORMATION			
ID	**Application Domain(s)**	**Name of KPI**	
KPI-AF-029	Railway	AI response time	
Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	07.02.2025	Roman Liessner (DB)	Creation of the document
1.0	03.03.2025	Roman Liessner (DB)	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	The Response Time KPI measures the time taken by the AI-assisted railway re-scheduling system to generate a new operational schedule in response to a disruption. This metric evaluates how quickly the system reacts to unexpected events, ensuring minimal delays and maintaining operational efficiency.
Objective(s)	This KPI contributes to evaluating Effectiveness of the AI-based assistant, as part of Task 4.1 evaluation objectives, and O2 main project objective: • To assess the speed of AI-assisted decision-making in railway operations. • To ensure rapid re-scheduling of trains in response to disturbances, minimizing the impact on passengers and freight. • To compare AI-assisted response times with traditional manual re-scheduling approaches.
Formula(s)	$ResponseTime = \frac{\sum[T_{proposal} - T_{detection}]}{n_{rescheduling}}$
Unit of Measurement	Time (minutes or seconds)

CALCULATION METHODOLOGY		
Baseline calculation methodology		
Step #	**Step description**	**Calculation**
1	Collect historical response times from manual re-scheduling processes.	Data Analysts
2	Establish a baseline response time benchmark based on past operational data.	Railway Control Center
KPI calculation methodology		
KPI step #	**Step description**	**Calculation**
1	Detect a disruption in the railway network.	Traffic Management System (TMS)
2	Record the time of disruption detection.	AI Monitoring System
3	Log the timestamp when the AI system proposes a new schedule.	AI System
4	Compute response time using the formula above.	Data Analysts
5	Compare AI-generated response times with historical manual benchmarks.	Railway Operators

Data collection			
Data ID	**Type**	**Source**	**Frequency**
$T_{detection}$	time of disruption detection	Scheduling	Each step (realtime)
$T_{proposal}$	timestamp when the AI system proposes a new schedule	Scheduling	Each step (realtime)
$n_{rescheduling}$	number of rescheduling events	Scheduling	Each step (realtime)

SOLUTION QUALITY

BASIC KPI INFORMATION			
ID	**Application Domain(s)**	**Name of KPI**	
KPI-CF-012	Power Grid	Carbon intensity	
Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	10.01.2025	B. LEMETAYER (RTE)	Creation of the document
1.0	03.03.2025	B. Lemetayer (RTE)	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	Carbon intensity estimates the overall carbon intensity of the action recommendation provided by the AI assistant to the human operator: goal of carbon intensity KPI is to measure how much the actions will directly contribute to greenhouse gases emission, by focusing on CO2 (which is unfortunately not the only greenhouse gas). It is calculated as the weighted averaged emission factor of generation variation, including: • Redispatching actions, • Curtailment actions. Details on source for emission factors: Methodology | Electricity Maps ipcc_wg3_ar5_annex-iii.pdf
Objective(s)	This KPI is calculated to estimate the relative performance compared to a baseline. The main difficulty of evaluating and assessing this KPIs lies in the difficulty to establish a proper deadline: • There is no history of human actions on the digital environments used for evaluation (since they are synthetic ones), • Comparison with KPI calculated on real grid's operations (TenneT or RTE) is not relevant since each grid has its own generation mix, and each TSO has its own operation policies (and own redispatching offers). This KPI contributes to evaluating Solution quality of the AI-based assistant, as part of Task 4.1 evaluation objectives, and O2 main project objective.
Formula(s)	See calculation steps
Unit of Measurement	kgCO2/MWh

CALCULATION METHODOLOGY		
Baseline calculation methodology		
Step #	**Step description**	**Module(s)**
1	Apply all KPI calculation steps and calculate the KPI for a baseline agent on the same environment: ideally this should be close to the human behavior	Digital environment KPI module
KPI calculation methodology		
KPI step #	**Step description**	**Module(s)**
1	Gather type of each generator (can be "solar", "wind", "hydro", "thermal" and "nuclear") in the environment Use *env.gen_type* in Grid2Op to get the vector of generators type for a given environment.	Digital environment KPI module

2	For each observation obs in Grid2Op, get amount of active power (P_{before} in MW) before curtailment of RES generation, use *obs.gen_p_before_curtail* vector in Grid2Op.	Digital environment
3	For each observation obs in Grid2Op, get amount of active power (P_{after} in MW) after curtailment of RES generation, use *obs.gen_p* vector in Grid2Op.	Digital environment
4	For each observation obs in Grid2Op, calculate the amount of energy variation due to curtailment for each RES generator g. Vector of energy variation due to curtailment of RES generation in MWh for each generator *g* over all observations is calculated as: $$E^{g}_{curtailment} = \sum_{obs} (P^{g,obs}_{after} - P^{g,obs}_{before}) \times duration_{step}$$ With $duration_{step}$ the duration of each step expressed in hours.	Digital environment KPI module
5	For each observation obs in Grid2Op, calculate the amount of energy variation due to redispatching for each generator g. Vector of energy variation due to redispatch in MWh for each generator *g* over all observations is calculated as: $$E^{g}_{redispatch} = \sum_{obs} P^{g,obs}_{redispatch} \times duration_{step}$$ With $duration_{step}$ the duration of each step expressed in hours.	Digital environment KPI module
6	For each observation obs in Grid2Op, calculate the total amount of energy variation for each generator g : $$E^{g} = E^{g}_{curtailment} + E^{g}_{redispatch}$$	Digital environment KPI module
7	Assign each generator its generation type i.	Digital environment KPI module
8	Sum the energy variation according to generation type, to get a vector with one value E_i in MWh per generation type i.	Digital environment KPI module
9	Gather emission factors (kgCO2/MWh) F_i for each generation type i.	Digital environment KPI module
10	Calculate the KPI as follows: $$CO_{2_{intensity}} = \frac{\sum E_i \times F_i}{\sum E_i}$$	Digital environment KPI module

Data collection

Data ID	*Type*	*Source*	*Frequency*
F_i	Emission factors	**Generation type** / **Emission factor (kgCO2/MWh)**: **Solar** 45 **Wind** 11 **Hydro** 24 **Thermal** 655[14] **Nuclear** 12	Once (no update)
obs	Observation	Grid2Op environment observation	Each episode step
P_{before}	Observation	Grid2Op environment observation property	Each episode step

[14] Calculated as the average of emission factors for coal (820 kgCO2/MWh) and gas (490 kgCO2/MWh)

P_{after}	Observation	Grid2Op environment observation property	Each episode step
$P_{redispatch}$	Observation	Grid2Op environment observation property	Each episode step

BASIC KPI INFORMATION			
ID	**Application Domain(s)**	**Name of KPI**	
KPI-TF-034	Power Grid	Topological action complexity	
Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	17.01.2025	B. LEMETAYER (RTE)	Creation of the document
1.0	03.03.2025	B. Lemetayer (RTE)	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	Topological action complexity KPI quantifies the topological utilization of the grid and gives insights into how many topological actions are utilized: performing too complex or too many topology actions can indeed navigate the grid into topologies that are either unknown or hard to recover from for operators.
Objective(s)	This KPI contributes to evaluating Solution quality of the AI-based assistant, as part of Task 4.1 evaluation objectives, and O2 main project objective.
Formula(s)	This KPI yields a vector with 6 values, that are calculated over all scenarios' steps: • The minimum, maximum and average number of topological actions performed by the AI assistant per timestamp, • The minimum, maximum and average share of modified buses per timestamp.
Unit of Measurement	Vector of 6 values expressed as: • Number (first 3 values), • Percent (decimal number between 0% and 100%, last 3 values).

CALCULATION METHODOLOGY		
Baseline calculation methodology		
Step #	**Step description**	**Calculation**
1	Apply all KPI calculation steps and calculate the KPI for a baseline agent on the same environment: ideally this should be close to the human behavior	Digital environment KPI module
KPI calculation methodology		
KPI step #	**Step description**	**Calculation**
1	Retrieve all actions taken by the evaluated agent for each observation obs in Grid2Op, stored in a collection noted as: $\{actions_{obs}\}$ This can be done for a given episode using *episode.actions* in Grid2Op, where episode is a *EpisodeData class* object.	Digital environment

2	For each $actions_{obs}$ element of the vector $\{actions_{obs}\}$ obtained at step 1, calculate the number of topological sub-actions it is composed of. This is done by adding the length of each of the following $actions_{obs}$ attributes in Grid2Op: • *set_bus*, only if the element is a line • *change_bus* • *line_or_set_bus* • *line_ex_set_bus* • *line_or_change_bus* • *line_ex_change_bus* • *line_set_status* • *line_change_status* For each $actions_{obs}$ element of the vector $\{actions_{obs}\}$, this step yields a number of topological sub-actions $n_{obs}^{topology}$, and thus the following vector: $\langle n_{obs}^{topology} \rangle$	Digital environment KPI module
3	Using vector $\langle n_{obs}^{topology} \rangle$ obtained at step 2, calculate the minimum number of topological actions performed by the AI assistant per timestamp as following: $min_{topology} = min_{obs}(\langle n_{obs}^{topology} \rangle)$	Digital environment KPI module
4	Using vector $\langle n_{obs}^{topology} \rangle$ obtained at step 2, calculate the maximum number of topological actions performed by the AI assistant per timestamp as following: $max_{topology} = max_{obs}(\langle n_{obs}^{topology} \rangle)$	Digital environment KPI module
5	Using vector $\langle n_{obs}^{topology} \rangle$ obtained at step 2, calculate the average number of topological actions performed by the AI assistant per timestamp as following: $avg_{topology} = avg_{obs}(\langle n_{obs}^{topology} \rangle)$	Digital environment KPI module
6	Get the vector of number of connected buses for each observation obs in Grid2Op, noted as: $\langle n_{obs}^{bus} \rangle$ Use for example *obs.get_energy_graph()* to get the network graph of a given observation in Grid2Op, and then: • either the number of distinct *global_bus_id* property of each node • or the number of distinct couple (*local_bus_id*, *sub_id*) properties of each node	Digital environment KPI module
7	Calculate the vector of change of connected buses for each observation obs in Grid2Op, noted as: $\langle delta_{obs+1}^{bus} \rangle = \langle n_{obs+1}^{bus} \rangle - \langle n_{obs}^{bus} \rangle$ First element of the vector is equal to 0	Digital environment KPI module
8	Get the total number of all possible buses of the environment, noted as: n_{bus} Multiply *env.n_sub* and *env.n_busbar_per_sub* properties of the environment in Grid2Op	Digital environment KPI module
9	Using vector $\langle delta_{obs}^{bus} \rangle$ obtained at step 7 and value n_{bus} obtained at step 8, calculate the minimum share of modified buses per timestamp as following: $min_{bus} = min_{obs}\left(\frac{\langle delta_{obs}^{bus} \rangle}{n_{bus}} \times 100\right)$	Digital environment KPI module

10	Using vector $\langle delta_{obs}^{bus} \rangle$ obtained at step 7 and value n_{bus} obtained at step 8, calculate the maximum share of modified buses per timestamp as following: $$max_{bus} = max_{obs}\left(\frac{\langle delta_{obs}^{bus} \rangle}{n_{bus}} \times 100\right)$$	Digital environment KPI module
11	Using vector $\langle delta_{obs}^{bus} \rangle$ obtained at step 7 and value n_{bus} obtained at step 8, calculate the average share of modified buses per timestamp as following: $$avg_{bus} = avg_{obs}\left(\frac{\langle delta_{obs}^{bus} \rangle}{n_{bus}} \times 100\right)$$	Digital environment KPI module
12	Calculate the KPI as follows: $$\begin{matrix} min_{topology} \\ \left\langle \begin{matrix} max_{topology} \\ avg_{topology} \\ min_{bus} \\ max_{bus} \\ avg_{bus} \end{matrix} \right. \end{matrix}$$	Digital environment KPI module

Data collection			
Data ID	**Type**	**Source**	**Frequency**
obs	Observation	Grid2Op environment observation	Each episode step
$actions_{obs}$	Actions	Grid2Op environment action	Each episode step

BASIC KPI INFORMATION			
ID	**Application Domain(s)**	**Name of KPI**	
KPI-OF-036	Power Grid	Operation score	
Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	17.01.2025	B. LEMETAYER (RTE)	Creation of the document
0.2	05.02.2025	B. LEMETAYER (RTE)	Addition of remaining energy to be supplied in case of backout
1.0	03.03.2025	B. Lemetayer (RTE)	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	The operation score KPI for operating a power grid includes the cost of a blackout, the cost of energy losses on the grid, and the cost of remedial actions.
Objective(s)	This KPI contributes to evaluating Solution quality of the AI-based assistant, as part of Task 4.1 evaluation objectives, and O2 main project objective. This KPI is linked with project's Long Term Expected Impacts (LTEI): • (LTEI1)KPI_{S-1}, 15%-20% reduction in renewable energy curtailment due to optimal exploration of network flexibility with AI (see "Sum of curtailed RES energy volumes" below) • (LTEI1)KPI_{S-2}, 20%-30% avoided electricity demand shedding (see "Sum of remaining energy to be supplied in case of blackout" below)
Formula(s)	This KPI yields a vector with 8 values per episode: • Number topological actions performed by the AI assistant, • Number of redispatching actions (including but not limited to storage) performed by the AI assistant, • Sum of redispatched energy volumes, • Sum of balanced energy volumes, *Note: this element is influenced by the actions implemented in the environment to compensate imbalances between loads and generations* • Number of RES curtailment actions performed by the AI assistant, *Such actions correspond to cases where the agent decreases generation from renewable energy sources (from what would be possible given the current weather)* • Sum of curtailed RES energy volumes, • Sum of energy losses (estimated as difference between active power values of generations and loads), • Sum of remaining energy to be supplied in case of blackout.
Unit of Measurement	Vector of 8 values expressed as: • Number, • Number, • Energy in MWh, • Energy in MWh, • Number, • Energy in MWh, • Energy in MWh, • Energy in MWh. These values are expressed as raw values and will be possibly normalized during the evaluation to get fixed range values.

CALCULATION METHODOLOGY		
Baseline calculation methodology		
Step #	**Step description**	**Calculation**

1	Apply all KPI calculation steps and calculate the KPI for a baseline agent on the same environment: ideally this should be close to the human behavior	Digital environment KPI module
KPI calculation methodology		
KPI step #	***Step description***	***Calculation***
1	Apply calculation steps 1, 2 of KPI Topological action complexity.	Digital environment KPI module
2	Using vector $\langle n_{obs}^{topology} \rangle$ obtained at step 1, calculate the total number of topological actions performed by the AI assistant as following: $$n_{topology} = sum_{obs}(\langle n_{obs}^{topology} \rangle)$$	Digital environment KPI module
3	For each $actions_{obs}$ element of the vector $\{actions_{obs}\}$ obtained at step 1, get the list of redispatching sub-actions it is composed of. This is done by getting the following $actions_{obs}$ attributes in Grid2Op: • *redispatch* • *storage_p* This yields the following vector: $$\langle \langle subaction^{redispatch} \rangle_{obs} \rangle$$	Digital environment KPI module
4	For each $\langle subaction^{redispatch} \rangle_{obs}$ element of the vector $\langle \langle subaction^{redispatch} \rangle_{obs} \rangle$ obtained at step 3, calculate the number of redispatching sub-actions it is composed of. This yields a number of redispatching sub-actions $n_{obs}^{redispatch}$, and thus the following vector: $$\langle n_{obs}^{redispatch} \rangle$$	Digital environment KPI module
5	Calculate the number of redispatching actions performed by the AI assistant as following: $$n_{redispatch} = sum_{obs}(\langle n_{obs}^{redispatch} \rangle)$$	Digital environment KPI module
6	Calculate the amount of energy variation due to redispatching for each generator g and for each observation obs in Grid2Op: get the amount of active power ($P_{redispatch}$ in MW) redispatched, use *obs.actual_dispatch* vector in Grid2Op Volume of energy variation due to redispatch in MWh over all observations is calculated as: $$E_{redispatch} = \sum_{g,obs} P_{redispatch}^{g,obs} \times duration_{step}$$ With $duration_{step}$ the duration of each step expressed in hours.	Digital environment KPI module
7	Calculate the amount of energy variation due to balancing for each generator g and for each observation obs in Grid2Op: get the amount of balancing power as the difference between • P_{target} in MW which represents the target redispatching asked by the agent), use *obs.target_dispatch* vector in Grid2Op • P_{target} in MW which represents the actual redispatching implemented in the environment, use *obs.actual_dispatch* Volume of energy variation due to balancing in MWh over all observations is calculated as: $$E_{balancing} = \sum_{g,obs} (P_{actual}^{g,obs} - P_{target}^{g,obs}) \times duration_{step}$$ With $duration_{step}$ the duration of each step expressed in hours. *Note: Such actions correspond to cases where the agent modifies (increase or decrease) the generator output values*	Digital environment KPI module

8	For each $actions_{obs}$ element of the vector $\{actions_{obs}\}$ obtained at step 1, get the list of RES curtailment sub-actions it is composed of. This is done by getting the *curtail* $actions_{obs}$ attribute in Grid2Op. This yields the following vector: $\langle\langle subaction^{curtailmentRES}\rangle_{obs}\rangle$	Digital environment KPI module
9	For each $\langle subaction^{curtailment}\rangle_{obs}$ element of the vector $\langle\langle subaction^{curtailment}\rangle_{obs}\rangle$ obtained at step 7, calculate the number of curtailment sub-actions it is composed of. This yields a number of curtailment sub-actions $n_{obs}^{redispatching}$, and thus the following vector: $\langle n_{obs}^{curtailementRES}\rangle$	Digital environment KPI module
10	Calculate the number of curtailment actions performed by the AI assistant as following: $n_{curtailmentRES} = sum_{obs}(\langle n_{obs}^{curtailmentRES}\rangle)$	Digital environment KPI module
11	Calculate the amount of energy variation due to curtailment for each RES generator g and for each observation obs in Grid2Op: • Get amount of active power (P_{before} in MW) before curtailment, use *obs.gen_p_before_curtail* vector in Grid2Op • Get amount of active power (P_{after} in MW) after curtailment, use *obs.gen_p* vector in Grid2Op Volume of energy variation due to curtailment in MWh over all observations is calculated as: $E_{curtailmentRES} = \sum_{g,obs}(P_{after}^{g,obs} - P_{before}^{g,obs}) \times duration_{step}$ With $duration_{step}$ the duration of each step expressed in hours.	Digital environment KPI module
12	Calculate the amount of power loss for each observation obs in Grid2Op: • Get amount of active power generation ($P_{generation}$ in MW) for each generator g, use *obs.gen_p* vector in Grid2Op • Get amount of active power load (P_{load} in MW) for each load l, use *obs.load_p* vector in Grid2Op Volume of remaining energy to be supplied in case of backout is calculated as: $E_{losses} = \sum_{obs}\left[\sum_{g} P_{generation}^{g,obs} - \sum_{l} P_{load}^{l,obs}\right] \times duration_{step}$ With $duration_{step}$ the duration of each step expressed in hours.	Digital environment KPI module
	Calculate the amount of remaining energy to be supplied in case of backout: this corresponds to the amount of active power load (P_{load} in MW) for all loads l in the observation obs_{last} of last step before the step where a blackout happens: a blackout is characterized by value “True” of the “done” property of the environment, obtained at each step (*env.step(action) function in grid2op*) Volume of energy losses in MWh over all observations is calculated as: $E_{blackout} = \sum_{l} P_{load}^{l,obs_{last}} \times duration_{step}$ With $duration_{step}$ the duration of each step expressed in hours. In case all steps of the episode are run without blackout, then $E_{blackout}$ is equal to 0.	Digital environment KPI module

13	Calculate the KPI as follows: $n_{topology}$ $n_{redispatch}$ $\left\langle E_{redispatch} \right.$ $\left. E_{balancing} \right\rangle$ $n_{curtailmentRES}$ $E_{curtailmentRES}$ E_{losses} $E_{blackout}$	Digital environment KPI module

Data collection			
Data ID	***Type***	***Source***	***Frequency***
obs	Observation	Grid2Op environment observation	Each episode step
$actions_{obs}$	Actions	Grid2Op environment action	Each episode step

BASIC KPI INFORMATION			
ID	**Application Domain(s)**	**Name of KPI**	
KPI-AS-068	Power Grid	Assistant adaptation to user preferences	
Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	30.01.2025	B. LEMETAYER (RTE)	Creation of the document
1.0	03.03.2025	B. Lemetayer (RTE)	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	Assistant adaptation to user preferences assesses how the AI assistant adapts to operator's choices and preferences. The assistant provides several recommendations which represent different trade-offs of different objectives, and the operator eventually makes one single choice. This KPI assume that an estimation of epistemic uncertainty is calculated for each action recommendation, which can be used later by the human to select the action in a multi-objective setting. This KPIs thus aims at measuring: • Whether the choice that the operator makes is in the set of recommendations proposed by the assistant, • How is the recommendation chosen by the operator ranked compared to the other ones, • Whether the recommendation chosen by the operator has a high epistemic uncertainty compared to the other recommendations.
Objective(s)	This KPI contributes to evaluating Solution quality of the AI-based assistant, as part of Task 4.1 evaluation objectives, and O2 main project objective.
Formula(s)	See calculation steps: for this KPI, raw values are given as lists to allow different possible summary calculations.
Unit of Measurement	Vector with 6 values without units, for each step: • the lowest epistemic uncertainty of recommendations • the highest epistemic uncertainty of recommendations • the epistemic uncertainty of the recommendation chosen by the operator • the rank of the recommendation chosen by the operator • the total number of proposed recommendations • whether the choice that the operator makes is in the set of recommendations proposed by the assistant

CALCULATION METHODOLOGY		
Baseline calculation methodology		
Step #	**Step description**	**Calculation**
1	Apply all KPI calculation steps and calculate the KPI for a baseline agent on the same environment: ideally this should be close to the human behavior	Recommendation module
KPI calculation methodology		
KPI step #	**Step description**	**Calculation**
1	For each step, collect the list of recommendations r proposed by the AI assistant to the operator, noted as follows: $\{\langle r \rangle_{step}\}$	Recommendation module

2	For each recommendation r, collect the epistemic uncertainty $uncertainty(r)$ expressed in percent (from 0% to 100%).	Recommendation module
3	For each recommendation r, collect the rank of this recommendation among all proposed recommendations $rank(r)$ expressed as a number (from 1 for the first proposed recommendation, to $n_{step}^{recommendations}$ where $n_{step}^{recommendations}$ is the total number of proposed recommendations at the same step)	Recommendation module
4	For each step, collect the recommendation chosen by the operator noted as: r_{step}^{choice}	Recommendation module
6	For each step, compute whether the choice that the operator makes is in the set of recommendations proposed by the assistant: • if r_{step}^{choice} is not present in $\langle r \rangle_{step}$, set the indicator $choice_{step}$ value to 1 • Else, set the indicator $choice_{step}$ value to 0.	Recommendation module
7	The KPIs is defined as a vector containing, for each step: • the lowest epistemic uncertainty of recommendations $min(\langle uncertainty(r) \rangle_{step})$ • the highest epistemic uncertainty of recommendations $max(\langle uncertainty(r) \rangle_{step})$ • the epistemic uncertainty $uncertainty(r_{step}^{choice})$ of the recommendation chosen by the operator • the rank $rank(r_{step}^{choice})$ of the recommendation chosen by the operator • the total number of proposed recommendations, $n_{step}^{recommendations}$ • whether the choice that the operator makes is in the set of recommendations proposed by the assistant, $choice_{step}$ This gives the following vector: $\left\langle \begin{matrix} min(\langle uncertainty(r) \rangle_{step}) \\ max(\langle uncertainty(r) \rangle_{step}) \\ uncertainty(r_{step}^{choice}) \\ rank(r_{step}^{choice}) \\ n_{step}^{recommendations} \\ choice_{step} \end{matrix} \right\rangle_{step}$	Recommendation module

Data collection			
Data ID	***Type***	***Source***	***Frequency***
r	Action recommendation	Recommendation module	Each episode step
$uncertainty(r)$	Epistemic uncertainty of an action recommendation	Recommendation module	Each episode step
$rank(r)$	Rank of an action recommendation	Recommendation module	Each episode step
$choice$	Action	Recommendation module	Each episode step

BASIC KPI INFORMATION			
ID	**Application Domain(s)**	**Name of KPI**	
KPI-NF-045	Railway Network	Network Impact Propagation (NCS)	
Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	07.02.2025	Roman Liessner (DB)	Creation of the document
1.0	03.03.2025	Roman Liessner (DB)	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	The Network Impact Propagation KPI measures how disruptions in one part of the railway network affect the overall system, including delay propagation and congestion spillover. This KPI helps evaluate the cascading effects of local disturbances and the efficiency of AI-assisted re-scheduling in mitigating these effects.
Objective(s)	This KPI contributes to evaluating Solution quality of the AI-based assistant, as part of Task 4.1 evaluation objectives, and O2 main project objective. • To assess the **ripple effects** of disruptions across the railway network. • To quantify how effectively AI-assisted re-scheduling contains and mitigates propagation of delays. • To support **decision-making** in optimizing re-scheduling strategies for network-wide efficiency.
Formula(s)	Percentage (%) of affected trains relative to the initial disruption: $NCS = \frac{n_{trains}^{affected}}{n_{trains}}$
Unit of Measurement	Percentage (%)

CALCULATION METHODOLOGY		
Baseline calculation methodology		
Step #	**Step description**	**Calculation**
1	Collect historical data on train delays and their propagation across the network.	Data Analysts
2	Establish network-wide delay benchmarks from past operational scenarios.	Railway Control Center
KPI calculation methodology		
KPI step #	**Step description**	**Calculation**
1	Detect an initial disruption (e.g., infrastructure failure, train delay).	Traffic Management System (TMS)
2	Monitor how the disruption propagates through the network over time.	AI Monitoring System
3	Calculate the number of affected trains (or Affected Network Nodes)	Data Analysts
4	Compute the Network Impact Propagation using the formula above.	Data Analysts
5	Compare AI-managed scenarios against historical propagation benchmarks.	Railway Operators

Data collection			
Data ID	**Type**	**Source**	**Frequency**
n_{trains}	Total number of trains (or Total Network Nodes)	Scheduling	End of episode

$n_{trains}^{affected}$	Number of trains affected (or Affected Network Nodes)	Scheduling	Each step (realtime)

SCALABILITY

BASIC KPI INFORMATION		
ID	**Application Domain(s)**	**Name of KPI**
KPI-AF-050	ATM, PowerGrid, Railway	AI-Agent Scalability Training

Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	17.01.2025	A.Castagna (enliteAI) Herke van Hoof (UvA)	Document Creation
0.2	28.01.2025	A.Castagna (enliteAI)	Refinement pass
1.0	03.03.2025	A.Castagna (enliteAI) Herke van Hoof (UvA)	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	AI-Agent Scalability Training measures the elapsed time required by an AI-agent to reach a predefined performance threshold. Time measured both as wallclock time (seconds) as well as steps or episodes according to the domain needs. The performance is defined by the native reward formulation defined by the digital environment or by domain experts. The time to threshold is measured across: i) Different instance complexities; ii) Different hardware availability. The performance threshold is set empirically and is defined by the cumulative reward formulation specific to the application domain. Note that the reward formulation used to train the agent may differ. For case (i), the type of hardware used should be logged to interpret the wallclock time measurements.
Objective(s)	This KPI contributes to evaluating Scalability of the AI-based assistant, as part of Task 4.1 evaluation objectives, and O2 main project objective.
Formula(s)	Time taken to achieve a specific performance level during the training phase of an AI-agent, considering varying instance complexities and hardware availability
Unit of Measurement	Steps or Episodes and wall-clock time

CALCULATION METHODOLOGY		
Baseline calculation methodology		
Step #	**Step description**	**Calculation**
1	To compare the scalability of the methods, a baseline is not required. However, a specific instance complexity and hardware setup can be used as a reference for assessing scalability	n.a.
KPI calculation methodology		
KPI step #	**Step description**	**Calculation**
1	Determine a use-case specific performance metric, e.g. environment native reward function or suitably chosen use-case specific metric.	Domain experts

2	Define instances complexity - Define a sequence of scenarios with increasing complexity, such as increasing area, increasing number of agents, and other domain-specific parameters.	Domain experts
3	Define hardware availability – Define a sequence of hardware setups to train an agent on a fixed environment instance. For example, varying computational resources and memory can be considered.	n.a.
4	Define the performance thresholds, using the metric derived from Step 1, for the instances defined at Step 2.	Domain experts
5	For each method and scenario defined in Step 2, an agent is initialized from scratch and trained until it achieves the desired performance level according to Step 4. The time, expressed as steps or episodes, elapsed between initialization and the reaching of the threshold is recorded.	Recommendation module, Simulation engine
6	For each method and hardware instance defined in Step 3, an agent is initialized from scratch and trained until it achieves the desired performance level according to Step 4. The wall-clock time elapsed between initialization and the reaching of the threshold is recorded.	Recommendation module, Simulation engine
7	The final KPI consists of the sets of measurements resulting from Step 5 and Step 6. It consists of two curves: one representing training time as a function of environment complexity, and the other representing training time as a function of hardware availability.	n.a.

Data collection			
Data ID	***Type***	***Source***	***Frequency***
$ConvergenceTime$	Time required by an agent to converge	Prediction module	Data is collected throughout the entire simulation until the agent converges. Convergence is defined as the achievement of the minimal performance level defined at Step 4. The data is collected multiple times to meet the conditions outlined in Step 5 and Step 6.

BASIC KPI INFORMATION			
ID	**Application Domain(s)**	**Name of KPI**	
KPI-AF-051	ATM, PowerGrid, Railway	Agent Scalability Testing	
Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	24.01.2025	A.Castagna (enliteAI) Herke van Hoof (UvA)	Document Creation
0.2	07.02.2025	Herke van Hoof (UvA)	Editing pass based on feedback
1.0	03.03.2025	A.Castagna (enliteAI) Herke van Hoof (UvA)	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	Compare multiple trained agents, RL-based or not, based on the average inference time to sample one or multiple actions while increasing the complexity of the scenario analysed. Complexity is a domain-relevant concept that must be defined.
Objective(s)	This KPI contributes to evaluating Scalability of the AI-based assistant, as part of Task 4.1 evaluation objectives, and O2 main project objective.
Formula(s)	Inference time and performance of the trained AI agents as a function of instance complexity on standardized hardware.
Unit of Measurement	Time to be measured in seconds Performance to be measured using the environment native reward function or a suitably chosen use-case specific metric. Complexity to be defined in a use-case specific way, e.g., using a sequence of pre-defined scenarios increasing in complexity, such as increasing area, number of vehicles, nodes in the network.

CALCULATION METHODOLOGY		
Baseline calculation methodology		
Step #	**Step description**	**Calculation**
1	The metric is most useful to compare the scalability of two proposed methods, in which case a baseline is not necessary. However, a baseline can be obtained by evaluating this KPI for a baseline agent.	Recommendation module, simulation engine
KPI calculation methodology		
KPI step #	**Step description**	**Calculation**
1	Determine a use-case specific performance metric, e.g. environment native reward function or suitably chosen use-case specific metric.	N.A.
2	Define a use-case specific sequence of scenarios with increasing complexity, such as increasing area, number of vehicles, nodes in the network.	N.A.
3	Define hardware setup – Define a standardized (possibly domain-specific) hardware setup to train an agent, to keep timing results comparable.	N.A.
3	For each method, a trained agent is loaded and interacts with the designed scenarios of increased complexity. For each interaction, the inference time is recorded, and the performance metric is calculated. These values are averaged for each complexity level to form the KPI metric. The final KPI consists of two curves (possibly visualized in a dual-axis figure): inference time and performance as a function of complexity.	Recommendation module, simulation engine

Data collection			
Data ID	**Type**	**Source**	**Frequency**
$InferenceTime$	Inference time	Recommendation module, simulation engine	The inference time is recorded for each addressed scenarios of increased complexity.
$Performance$	Performance	Recommendation module, simulation engine	The performance is recorded for each method and addressed scenarios of increased complexity.

SAFETY AND ROBUSTNESS

This annex details all evaluation protocols corresponding to §4 - Safety and robustness, and Robustness, Resilience, Reliability objectives.

Page intentionally left blank.

ROBUSTNESS

BASIC KPI INFORMATION		
ID	**Application Domain(s)**	**Name of KPI**
KPI-RS-058	ATM, Power Grid, Railway	Robustness to operator input

Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	24.01.2025	Herke van Hoof (UvA)	Initial proposal
0.2	27.01.2025	Julia Usher (FNHW)	Review and addition for T3.4
0.3	07.02.2025	Herke van Hoof (UvA)	Editing pass based on feedback
1.0	03.03.2025	Herke van Hoof (UvA)	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	The KPI should measure or evaluate how the trained agent behaves in terms of robustness if, during the decision-making process where a human operator makes the final decisions, a human operator occasionally intervenes and significantly overrides the autonomous decisions of the trained agent. For agents trained using machine learning methods, this can cause an offset between the type of states encountered in the training data and during deployment, especially for agents trained using reinforcement learning or similar methods where the agent itself decides which actions to execute. As a consequence of this offset, the agent might make poorer decisions if the human operator does not always follow the proposed actions of the agents. To measure how sensitive the agent is to such offsets, this KPI proposes to use a “simulated operator” that does not fully follow the course of actions suggested by the agents,and instead overwrites certain action variables set by the agents in a fraction of time steps.
Objective(s)	Overall, this KPI contributes to evaluating Robustness of the AI-based assistant when dealing with real-world conditions, as part of Task 4.2 evaluation objectives, and O4 main project objective. The KPI is related to Tasks 3.1 and 3.3. Specifically, it is related to goal (4) of Task 3.1 (“Analysis of the impact of human intervention in the decision process on AI agents developed and trained towards fully autonomous behavior”), goal (1) of Task 3.3 (“Develop and expand order-agnostic network architectures adapted to the RL setting to use human-data or human-like perturbations and ensure the system can also make good decisions in the context where actions are partially chosen by the human partner”) and goal (2) of Task 3.4 (“Detect risks early on and potentially inform human supervisors, e.g. relinquish control to a human supervisor or transition into “safety mode” when necessary”).
Formula(s)	N.A.
Unit of Measurement	Environment reward, or the unit of measurement of a suitably-chosen use-case specific metric.

CALCULATION METHODOLOGY		
Baseline calculation methodology		
Step #	**Step description**	**Calculation**

1	Determine the performance (native environment reward, or use-case specific metric) of a baseline agent on the same environment, including the same perturbations by a simulated operator as detailed in step #1 of KPI calculation.	Recommendation module, simulation engine
2	Additional point of comparison can be obtained by determining the performance (in the same manner as under #1) of the primary agent when 0% of actions are chosen by the simulated operator.	Recommendation module, simulation engine
KPI calculation methodology		
KPI step #	***Step description***	***Calculation***
1	For each use-case, a domain specific choice is made regarding to define a simulated operator: • Percentage of the time and number of action variables chosen by the simulated operator. • Strategy to be followed by the simulated operator (e.g., completely at random, following a pre-defined heuristic, or following a strategy that attempts to simulate decisions in a logged dataset of real human inputs). Then, at each time step, with the chosen probability, the simulated operator modifies the suggestion made by the agent (if not, the suggestion is applied to the environment unchanged). In cases where the suggestion by the agent is modified, the simulated operator changes the chosen number of action variables following the chosen strategy. To reduce variance in comparisons using a randomized agent, the random seed of the simulated operator might be fixed to a set value for each evaluation scenario	N.A.
2	The performance of the primary AI agent (e.g., environment native reward function, or another metric defined for each use-case) is then measured in the presence of these deviations.	Recommendation module, simulation engine, digital environment

Data collection			
Data ID	***Type***	***Source***	***Frequency***
$performance$	Performance of the AI agent (e.g., environment native reward function, or another metric defined for each use-case)	Digital environment (or other appropriate module)	Episode run

BASIC KPI INFORMATION			
ID	**Application Domain(s)**	**Name of KPI**	
KPI-DF-069	ATM, PowerGrid, Railway	Drop-off in reward	
Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	27.01.2025	T. Tjhay (INESC TEC)	Creation of the document
1.0	03.03.2025	T. Tjhay (INESC TEC)	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	Drop-off in reward calculates difference in reward between situation with perfect information and imperfect information either through natural malfunctions while measuring data or through intentional perturbations by an attacker.
Objective(s)	This KPI contributes to evaluating Robustness of the AI-based assistant, as part of Task 4.2 evaluation objectives, and O4 main project objective.
Formula(s)	$\sum_{k}^{K} R_k^{unperturb} - R_k^{perturb}$ Where $R_k^{perturb}$ and $R_k^{unperturb}$ is the reward obtained in step *k* in the environment with and without perturbations, respectively.
Unit of Measurement	Same unit as reward or percentage of reward with perfect information

CALCULATION METHODOLOGY			
Baseline calculation methodology			
Step #	**Step description**		**Calculation**
1	Apply all KPI calculation steps and calculate the KPI for a baseline AI agent on the same environment		Digital environment, baseline AI agent
KPI calculation methodology			
KPI step #	**Step description**		**Calculation**
1	Run an episode of the digital environment with an AI agent that has perfect information until a game over or the maximum number of steps is reached		Digital environment, AI agent
2	Run the episode under the same circumstances but include a perturbation agent		Digital environment, AI agent, perturbation agent
3	Sum up the reward obtained in each step for both episodes and compute the difference		Digital environment
Data collection			
Data ID	**Type**	**Source**	**Frequency**
reward	Observation	Digital environment	Each step

BASIC KPI INFORMATION			
ID	**Application Domain(s)**	**Name of KPI**	
KPI-FF-070	ATM, PowerGrid, Railway	Frequency changed output AI agent	
Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	27.01.2025	T. Tjhay (INESC TEC)	Creation of the document
1.0	03.03.2025	T. Tjhay (INESC TEC)	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	Frequency changed output AI agent calculates the number of times the output of the AI agent (i.e. the action the agent chooses) is changed due to perturbations
Objective(s)	This KPI contributes to evaluating Robustness of the AI-based assistant, as part of Task 4.2 evaluation objectives, and O4 main project objective.
Formula(s)	Count number of times (steps) the action the AI agent chooses based on unperturbed and perturbed input is different
Unit of Measurement	None (number)

CALCULATION METHODOLOGY		
Baseline calculation methodology		
Step #	**Step description**	**Calculation**
1	Apply all KPI calculation steps and calculate the KPI for a baseline AI agent on the same environment	Digital environment, baseline AI agent
KPI calculation methodology		
KPI step #	**Step description**	**Calculation**
1	Run an episode in the digital environment and feed the AI agent both the unperturbed and perturbed input during each step	Digital environment, AI agent, perturbation agent
2	Compare the actions the agent chooses and count how many times the actions are different for the two inputs	Digital environment, AI agent
3	Use the action for the perturbed input to advance to the next step in the environment	Digital environment, AI agent

Data collection			
Data ID	**Type**	**Source**	**Frequency**
$act^{unpert.}$	Action chosen by AI-based agent without any perturbations to input	Output AI agent	Each step
$act^{pert.}$	Action chosen by AI-based agent with perturbed input	Output AI agent	Each step

BASIC KPI INFORMATION		
ID	**Application Domain(s)**	**Name of KPI**
KPI-SF-071	ATM, PowerGrid, Railway	Severity of changed output AI agent

Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	27.01.2025	T. Tjhay (INESC TEC)	Creation of the document
1.0	03.03.2025	T. Tjhay (INESC TEC)	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	Severity of changed output AI agent KPI measures similarity of the action chosen by AI agent based on a perturbed input to the action chosen with perfect information. Average pre-defined similarity score per changed action indicating how different the new action is from the original one.
Objective(s)	This KPI contributes to evaluating Robustness of the AI-based assistant, as part of Task 4.2 evaluation objectives, and O4 main project objective.
Formula(s)	A similarity score is defined for every pair of actions in the action space of the AI agent using the following formulas: $$C^{a^1,a^2} = \frac{1}{2}\left(\left|c^{a^1,a^2}\right| + \frac{\left|\hat{c}^{a^1,a^2}\right|}{2}\right)\left(\frac{1}{\left|c^{a^1}\right|} + \frac{1}{\left|c^{a^2}\right|}\right) \quad (1)$$ $$V^{a^1,a^2} = \frac{1}{2}\left(\frac{\left|v^{a^1,a^2}\right|}{\left|v^{a^1}\right|} + \frac{\left|v^{a^1,a^2}\right|}{\left|v^{a^2}\right|}\right) \quad (2)$$ $$\frac{V^{a^1,a^2} + C^{a^1,a^2}}{2} \quad (3)$$ The similarity score can be computed as the average of two parts (Eq. 3), a part that evaluates the exact same changes (Eq. 1) and one that accounts for changes to the same substation/node in the network (Eq. 2). In these equations, the set c^a consists of the changes that are made by action a, in the power grid case this could be setting the origin of powerline 1 at busbar 1, and $c^{a^1,a^2} = c^{a^1} \cap c^{a^2}$ is the set of changes that are made in both actions a^1 and a^2. Additionally, $\hat{c}^{a^1,a^2}$ is the set of changes that are almost the same, e.g., an action setting the extremity of line 4 at bus 2 and an action setting it at bus 1. Similarly in Eq. 6, v^a is the set of all substations/nodes affected by action a, and v^{a^1,a^2} is the set of substations/nodes affected by both a^1 and a^2. Metric can be computed by summing similarity scores for each step where the action is changed due to perturbation in an episode and dividing it by the number of the number of steps with an action change (KPI-FT-070) to get the average similarity score per action change
Unit of Measurement	Average similarity score per action change

CALCULATION METHODOLOGY		
Baseline calculation methodology		
Step #	**Step description**	**Calculation**
1	Apply all KPI calculation steps and calculate the KPI for a baseline AI agent on the same environment	Digital environment, baseline AI agent
KPI calculation methodology		
KPI step #	**Step description**	**Calculation**

1	Define a similarity score for each pair of actions in the action space of the AI agent	Digital environment, AI agent
2	Run an episode in the digital environment and feed the AI agent both the unperturbed and perturbed input during each step	Digital environment, AI agent, perturbation agent
3	Compare the actions the agent chooses and if the two are different add the similarity score of the two actions to the total similarity score for the episode	Digital environment, AI agent
4	Divide the total similarity score by the number of times the action was changed (KPI-FT-070)	Digital environment, AI agent

Data collection			
Data ID	***Type***	***Source***	***Frequency***
$act^{unpert.}$	Action chosen by AI-based agent without any perturbations to input	Output AI agent	Each step
$act^{pert.}$	Action chosen by AI-based agent with perturbed input	Output AI agent	Each step

+ BASIC KPI INFORMATION			
ID	**Application Domain(s)**	**Name of KPI**	
KPI-SF-072	ATM, PowerGrid, Railway	Steps survived with perturbations	
Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	27.01.2025	T. Tjhay (INESC TEC)	Creation of the document
1.0	03.03.2025	T. Tjhay (INESC TEC)	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	Steps survived with perturbations KPI calculates the number of steps the AI agent is able to survive in environment with perturbation agent
Objective(s)	This KPI contributes to evaluating Robustness of the AI-based assistant, as part of Task 4.2 evaluation objectives, and O4 main project objective.
Formula(s)	Count the number of steps survived in the digital environment without a game over
Unit of Measurement	Number of steps

CALCULATION METHODOLOGY		
Baseline calculation methodology		
Step #	**Step description**	**Calculation**
1	Run an episode of the digital environment with an AI agent that has perfect information until a game over or the maximum number of steps is reached	Digital environment, AI agent
2	Get total number of steps survived without game over	Digital environment
KPI calculation methodology		
KPI step #	**Step description**	**Calculation**
1	Run an episode of the digital environment with an AI agent and a perturbation agent until a game over or the maximum number of steps is reached	Digital environment, AI agent, perturbation agent
2	Get total number of steps survived without game over	Digital environment

Data collection			
Data ID	**Type**	**Source**	**Frequency**
n_steps	Number of steps survived without game over	Digital environment	Each episode

BASIC KPI INFORMATION			
ID	**Application Domain(s)**	**Name of KPI**	
KPI-RF-078	ATM, PowerGrid, Railway	Reward per action	
Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	27.01.2025	T. Tjhay (INESC TEC)	Creation of the document
1.0	03.03.2025	T. Tjhay (INESC TEC)	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	Reward per action KPI calculates average reward obtained for each action performed by the AI agent
Objective(s)	This KPI contributes to evaluating Robustness of the AI-based assistant, as part of Task 4.2 evaluation objectives, and O4 main project objective.
Formula(s)	$\frac{\sum_{k=0}^{K} R_k}{\sum_{k=0}^{K} \mathbf{1}_{a_k \neq a^{\emptyset}}}$ Here, R_k is the reward obtained in step *k* and $\mathbf{1}_{a_k \neq a^{\emptyset}}$ is an indicator function that returns 1 if the AI-based agent performs an action in step *k* and 0 otherwise.
Unit of Measurement	Same unit as reward

CALCULATION METHODOLOGY		
Baseline calculation methodology		
Step #	**Step description**	**Calculation**
1	Run an episode of the digital environment with an AI agent that has perfect information until a game over or the maximum number of steps is reached	Digital environment, AI agent
2	Sum up the reward obtained in each step and divide it by the number of times the AI agent took an action	Digital environment, AI agent
KPI calculation methodology		
KPI step #	**Step description**	**Calculation**
1	Run an episode of the digital environment with an AI agent and a perturbation agent until a game over or the maximum number of steps is reached	Digital environment, AI agent, perturbation agent
2	Sum up the reward obtained in each step and divide it by the number of times the AI agent took an action	Digital environment, AI agent

Data collection			
Data ID	**Type**	**Source**	**Frequency**
reward	Observation	Digital environment	Each step
act	Action chosen by AI-based agent based on input	Output AI agent	Each step

BASIC KPI INFORMATION			
ID	**Application Domain(s)**	**Name of KPI**	
KPI-VF-073	ATM, PowerGrid, Railway	Vulnerability to perturbation	
Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	27.01.2025	T. Tjhay (INESC TEC)	Creation of the document
1.0	03.03.2025	T. Tjhay (INESC TEC)	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	Vulnerability to perturbation KPI measures vulnerability of specific value in observed state to perturbations, i.e. how likely it is that perturbing the value will result in a change in action chosen by the AI agent
Objective(s)	This KPI contributes to evaluating Robustness of the AI-based assistant, as part of Task 4.2 evaluation objectives, and O4 main project objective.
Formula(s)	For a given value in the state space of the environment, $state_i$, compute the mean, μ, and standard deviation, σ, of the perturbation applied to the value in each step and define a threshold that determines whether the value is significantly perturbed in a step or not as $\mu \pm \sigma$. The metric can then be computed using Eq. 1. $$\frac{\sum_{k=0}^{K} \mathbf{1}_{a_k^{adv} \neq a_k} \mathbf{1}_{s_i k \text{ is perturbed}}}{\sum_{k=0}^{K} \mathbf{1}_{s_i k \text{ is perturbed}}} \quad (1)$$ Count the number of times the value is perturbed significantly in the episode, $\sum_{k=0}^{K} \mathbf{1}_{s_i k \text{ is perturbed}}$, and the number of times the value is perturbed in a step where the action of the AI agent is changed, $\sum_{k=0}^{K} \mathbf{1}_{a_k^{adv} \neq a_k} \mathbf{1}_{s_i k \text{ is perturbed}}$. Then divide the latter by the former to compute the metric for x_i.
Unit of Measurement	Proportion of times perturbing the value resulted in a changed action

CALCULATION METHODOLOGY		
Baseline calculation methodology		
Step #	**Step description**	**Calculation**
1	Apply all KPI calculation steps and calculate the KPI for a baseline AI agent on the same environment	Digital environment, baseline AI agent, perturbation agent
KPI calculation methodology		
KPI step #	**Step description**	**Calculation**
1	Run an episode in the digital environment and feed the AI agent both the unperturbed and perturbed input during each step, while keeping track of the perturbation performed in each step	Digital environment, AI agent, perturbation agent
2	For a given value, get the mean and standard deviation of the perturbations of the value in each step and define a threshold indicating whether the value was significantly perturbed in a step.	Digital environment, AI agent
3	Count the number of steps where the value was perturbed significantly.	Digital environment, perturbation agent
4	Count the number of steps where the value was perturbed significantly, and the action of the AI agent was changed.	Digital environment, perturbation agent, AI agent

5	Divide the number from step 4 by the number from step 3	n.a.
6	Repeat for every value in the state space of the environment	Digital environment, perturbation agent, AI agent

Data collection			
Data ID	**Type**	**Source**	**Frequency**
$act^{unpert.}$	Action chosen by AI-based agent without any perturbations to input	Output AI agent	Each step
$act^{pert.}$	Action chosen by AI-based agent with perturbed input	Output AI agent	Each step
perturb	Perturbation applied to the input of the AI-based agent	Output perturbation agent	Each step
state	Observed state	Digital environment	Each step

BASIC KPI INFORMATION			
ID	**Application Domain(s)**	**Name of KPI**	
KPI-EF-087	ATM, Power Grid, Railway	Explainability Faithfulness	
Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	25.02.2025	Mohamed Hassouna (Fraunhofer)	Initial Draft
1.0	03.03.2025	Mohamed Hassouna (Fraunhofer)	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	The *Faithfulness* KPI assesses whether the feature importance scores provided by an explanation method accurately reflect the model's decision-making process. It systematically removes or alters features and measures the impact on the model's predictions. The assumption is that if a feature is truly important, removing or altering it should significantly affect the model's output.
Objective(s)	This KPI ensures that AI-driven explanations remain reliable and aligned with the actual decision-making process of the model. It helps evaluate interpretability methods in AI systems used in critical applications. This KPI contributes to evaluating AI trustworthiness, acceptability and trust of the AI-based assistant, as part of Task 4.3 evaluation objectives, and O4 main project objective.
Formula(s)	A dataset with model/agent predictions and corresponding explanation e.g. importance scores shall be collected. $$FE = \sum_{i=1}^{N} \lvert f(x_i) - f(x_i^{\setminus j}) \rvert$$ where: • FE is the Faithfulness Estimate • $f(x_i)$ is the model prediction for the original input x_i • $f(x_i^{\setminus j})$ is the model prediction when feature j is removed, masked, or replaced • N is the total number of evaluated samples
Unit of Measurement	Change in model confidence score (e.g., probability difference), Normalized score indicating faithfulness

CALCULATION METHODOLOGY		
Baseline calculation methodology		
Step #	**Step description**	**Calculation**
1	Identify dataset samples and their corresponding feature importance scores	KPI module
2	Apply systematic feature removal (e.g., masking, perturbation, mean imputation) i.e. modify features $x_i^{\setminus j}$	KPI module
KPI calculation methodology		
KPI step #	**Step description**	**Calculation**
1	Compute the faithfulness estimate for each feature removal	KPI Module
2	Aggregate results over multiple runs	KPI module
3	• Compare FE values with feature importance scores. • Compute correlation (e.g., Pearson or Spearman) between FE and feature importance scores from the explanation method	KPI module
4	Normalize and interpret results	KPI module

Data collection			
Data ID	**Type**	**Source**	**Frequency**
$f(x)$	Model predictions	AI model input/output dataset	Every inference
$\Phi(f,x)$	Feature importance scores	Explanation method output	Every inference

BASIC KPI INFORMATION			
ID	**Application Domain(s)**	**Name of KPI**	
KPI-EF-086	ATM, Power Grid, Railway	Explainability Robustness	
Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	25.02.2025	Mohamed Hassouna (Fraunhofer)	Initial Draft
1.0	03.03.2025	Mohamed Hassouna (Fraunhofer)	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	The *Explainability Robustness KPI* evaluates the stability of explanations against small input perturbations, assuming the model's output remains relatively unchanged. A robust explanation should not fluctuate significantly when the input is slightly modified. The *Average Sensitivity Metric* quantifies this stability by applying small perturbations to the input data and measuring how much the explanation changes. Since computing sensitivity over all possible perturbations is impractical, Monte Carlo sampling is used to estimate these variations efficiently.
Objective(s)	This KPI ensures that AI-driven explanations remain reliable and aligned with the actual decision-making process of the model. It helps evaluate interpretability methods in AI systems used in critical applications. This KPI contributes to evaluating AI trustworthiness, acceptability and trust of the AI-based assistant, as part of Task 4.3 evaluation objectives, and O4 main project objective.
Formula(s)	A dataset with model/agent predictions shall be created or re-used. $$S = E_{\delta \sim \mathbb{D}}[\| \Phi(f,x) - \Phi(f,x+\delta)\|^p]^{1/p}$$ Where: • S is the Average Sensitivity • $\Phi(f,x)$ represents the explanation for the original input x • $\Phi(f,x+\delta)$ represents the explanation for the perturbed input $x+\delta$ • $\delta \sim \mathbb{D}$ is the small perturbation sampled from a predefined distribution • $\|\cdot\|^p$ denotes the p-norm (e.g., L1 or L2 distance) measuring explanation differences
Unit of Measurement	Change in explanation values (e.g., L1 or L2 norm difference), Normalized score indicating robustness

CALCULATION METHODOLOGY		
Baseline calculation methodology		
Step #	**Step description**	**Calculation**
1	Identify dataset samples and their corresponding explanation e.g. feature importance scores	KPI module
2	Apply small random perturbations to inputs	KPI module
3	Recalculate explanations for perturbed input	KPI Module
KPI calculation methodology		
KPI step #	**Step description**	**Calculation**
1	Compute explanation differences, i.e. sensitivity estimate for each sample: $$\| \Phi(f,x) - \Phi(f,x+\delta)\|^p$$	KPI Module

2	Aggregate results over multiple runs: $S = E_{\delta \sim \mathbb{D}}[\|\| \Phi(f,x) - \Phi(f,x+\delta)\|\|^p]^{1/p}$	KPI module
4	Normalize and interpret results	KPI module

Data collection			
Data ID	***Type***	***Source***	***Frequency***
$f(x)$	Model predictions	AI model input/output dataset	Every inference
$\Phi(f,x)$	Feature importance scores	Explanation method output	Every inference

RESILIENCE

BASIC KPI INFORMATION		
ID	**Application Domain(s)**	**Name of KPI**
KPI-AF-074	ATM, PowerGrid, Railway	Area between reward curves

Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	27.01.2025	T. Tjhay (INESC TEC)	Creation of the document
1.0	03.03.2025	T. Tjhay (INESC TEC)	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	Area between reward curves calculates area between the curve corresponding to the reward obtained in each step in an environment where the AI agent has perfect information and the curve for an environment where the agent's input is perturbed
Objective(s)	This KPI contributes to evaluating Resilience of the AI-based assistant, as part of Task 4.2 evaluation objectives, and O4 main project objective.
Formula(s)	Use the trapezoidal rule for numerical integration (Eq. 1) to compute the area underneath the two curves and subtract the area of the perturbed situation from the area of the one with perfect information. $$\int_1^K \Delta R_k dk \approx \frac{\Delta R_K + \Delta R_1}{2} + \sum_{i=2}^{K-1} \Delta R_i \qquad (1)$$ In this equation ΔR_k is the difference between the reward obtained in step *k* in the perturbed and unperturbed environment.
Unit of Measurement	None (cumulative reward)

CALCULATION METHODOLOGY		
Baseline calculation methodology		
Step #	**Step description**	**Calculation**
1	Apply all KPI calculation steps and calculate the KPI for a baseline AI agent on the same environment	Digital environment, baseline AI agent
KPI calculation methodology		
KPI step #	**Step description**	**Calculation**
1	Run an episode of the digital environment with an AI agent until a game over or the maximum number of steps is reached	Digital environment, AI agent
2	Run the same episode again but this time with a perturbation agent	Digital environment, AI agent, perturbation agent
3	Compute the difference between the reward with and without perturbations, ΔR_k, for every step k	Digital environment
4	Use the trapezoidal rule (Eq. 1) to approximate the area between the reward curves	KPI module

Data collection			
Data ID	**Type**	**Source**	**Frequency**
$reward^{unpert}$	Observation	Digital environment	Each step
$reward^{pert}$	Observation	Digital environment	Each step

BASIC KPI INFORMATION			
ID	**Application Domain(s)**	**Name of KPI**	
KPI-DF-075	ATM, PowerGrid, Railway	Degradation time	
Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	27.01.2025	T. Tjhay (INESC TEC)	Creation of the document
1.0	03.03.2025	T. Tjhay (INESC TEC)	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	Number of steps/episodes until reward reaches its lowest point after introducing perturbations to the input of the AI agent
Objective(s)	This KPI contributes to evaluating Resilience of the AI-based assistant, as part of Task 4.2 evaluation objectives, and O4 main project objective.
Formula(s)	Find the step/episode where the reward is the lowest using Eq.1, where R_{hk} is the reward obtained in step *k* of episode *h,* and get the difference to the step/episode the perturbations were introduced, h^p, i.e. $h^{min} - h^p$. $$\mathrm{h^{min}} = \mathrm{argmin}_{\mathrm{h^p} \le \mathrm{h} \le \mathrm{H}} \{ \sum_{\mathrm{k=1}}^{\mathrm{K}} R_{hk} \} \quad (1)$$
Unit of Measurement	Number of steps/episodes

CALCULATION METHODOLOGY		
Baseline calculation methodology		
Step #	**Step description**	**Module(s)**
1	Apply all KPI calculation steps and calculate the KPI for a baseline AI agent on the same environment	Digital environment, baseline AI agent
KPI calculation methodology (Training phase)		
KPI step #	**Step description**	**Module(s)**
1	Train AI agent for number of episodes while keeping track of average reward per step in each episode.	Digital environment, AI agent
2	Train AI agent again under same circumstances but include a perturbation agent after certain number of episodes, h^p.	Digital environment, AI agent, perturbation agent
3	Find episode where difference between unperturbed and perturbed reward per step is largest, h^{min}.	Digital environment
4	Get number of episodes between h^p and h^{min}.	Digital environment KPI module
KPI calculation methodology (Testing phase)		
KPI step #	**Step description**	**Module(s)**
1	Run an episode of the digital environment with an AI agent until a game over or the maximum number of steps is reached	Digital environment, AI agent
2	Run the episode again under same circumstances but include a perturbation agent	Digital environment, AI agent, perturbation agent

3	Find step where the reward begins to differ between the two episodes, h^p, and the step where the difference between unperturbed and perturbed reward per step is largest, h^{min}.	Digital environment
4	Get number of steps between h^p and h^{min}.	Digital environment KPI module

Data collection			
Data ID	***Type***	***Source***	***Frequency***
reward	Observation	Digital environment	Each step

BASIC KPI INFORMATION			
ID	**Application Domain(s)**	**Name of KPI**	
KPI-RF-076	ATM, PowerGrid, Railway	Restorative time	
Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	27.01.2025	T. Tjhay (INESC TEC)	Creation of the document
1.0	03.03.2025	T. Tjhay (INESC TEC)	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	Number of steps/episodes until reward recovers to its highest point after reaching the lowest point after introducing perturbations to the input of the AI agent
Objective(s)	This KPI contributes to evaluating Resilience of the AI-based assistant, as part of Task 4.2 evaluation objectives, and O4 main project objective.
Formula(s)	Find the step/episode where the reward is the highest, h^{max} , and the one with the lowest reward from KPI-DT-074, h^{min} using Eq. 2 and Eq. 1, respectively, where R_{hk} is the reward obtained in step *k* of episode *h*. $$h^{min} = argmin_{h^p \le h \le H}\{\sum_{k=1}^{K} R_{hk}\} \quad (1)$$ $$h^{max} = argmax_{h^{min} \le h \le H}\{\sum_{k=1}^{K} R_{hk}\} \quad (2)$$ Subtract them from each other, i.e. $h^{max} - h^{min}$, to get the restorative time
Unit of Measurement	Number of steps/episodes

CALCULATION METHODOLOGY		
Baseline calculation methodology		
Step #	**Step description**	**Module(s)**
1	Apply all KPI calculation steps and calculate the KPI for a baseline AI agent on the same environment	Digital environment, baseline AI agent
KPI calculation methodology (Training phase)		
KPI step #	**Step description**	**Module(s)**
1	Train AI agent for number of episodes while keeping track of average reward per step in each episode.	Digital environment, AI agent
2	Train AI agent again under same circumstances but include a perturbation agent after certain number of episodes, h^p.	Digital environment, AI agent, perturbation agent
3	Find episode where difference between unperturbed and perturbed reward per step is largest, h^{min}, and the episode with the highest reward per step after h^{min}, h^{max}.	Digital environment
4	Get number of episodes between h^{min} and h^{max}.	Digital environment KPI module
KPI calculation methodology (Testing phase)		
KPI step #	**Step description**	**Module(s)**
1	Run an episode of the digital environment with an AI agent until a game over or the maximum number of steps is reached	Digital environment, AI agent

2	Run the episode again under same circumstances but include a perturbation agent	Digital environment, AI agent, perturbation agent
3	Find the step where the difference between unperturbed and perturbed reward per step is largest, h^{min}, and the step where the difference in reward is the smallest after h^{min}, h^{max}	Digital environment
4	Get number of steps between h^{min} and h^{max}.	Digital environment KPI module

Data collection			
Data ID	***Type***	***Source***	***Frequency***
reward	Observation	Digital environment	Each step

BASIC KPI INFORMATION			
ID	**Application Domain(s)**	**Name of KPI**	
KPI-SF-077	ATM, PowerGrid, Railway	Similarity state to unperturbed situation	
Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	27.01.2025	T. Tjhay (INESC TEC)	Creation of the document
1.0	03.03.2025	T. Tjhay (INESC TEC)	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	Similarity state to unperturbed situation KPI measures similarity of the state in an environment where AI agent's input is perturbed to the state in the same context of an environment with perfect information
Objective(s)	This KPI contributes to evaluating Resilience of the AI-based assistant, as part of Task 4.2 evaluation objectives, and O4 main project objective.
Formula(s)	Choose a metric to measure the similarity between states, e.g. cosine similarity or Euclidean distance between vector representations of the state, and compute similarity between the state in each step of environment with perfect information and one with perturbed input. Plot curve of similarity in each step and evaluate using KPI-AT-073, KPI-DT-074 and KPI-RT-075. As an example, the cosine similarity is chosen as shown in Eq. 1: $$\frac{\sum_{i=0}^{\lvert s\rvert} s_i^{adv} s_i}{\sqrt{\sum_{i=0}^{\lvert s\rvert} (s_i^{adv})^2}\sqrt{\sum_{i=0}^{\lvert s\rvert} (s_i)^2}} \quad (1)$$
Unit of Measurement	none

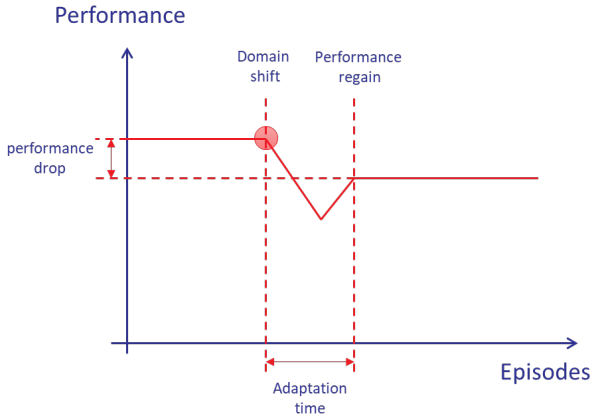
CALCULATION METHODOLOGY		
Baseline calculation methodology		
Step #	**Step description**	**Calculation**
1	Apply all KPI calculation steps and calculate the KPI for a baseline AI agent on the same environment	Digital environment, baseline AI agent
KPI calculation methodology		
KPI step #	**Step description**	**Calculation**
1	Run an episode of the digital environment with an AI agent until a game over or the maximum number of steps is reached	Digital environment, AI agent
2	Run the episode under the same circumstances but include a perturbation agent	Digital environment, AI agent, perturbation agent
3	For each step calculate the similarity between the state in the environment with the perturbation agent and the one without using the chosen metric	Digital environment
4	Evaluate KPI-AT-073, KPI-DT-074 and KPI-RT-075 for the similarity in each step	See relevant KPIs

Data collection			
Data ID	**Type**	**Source**	**Frequency**
state	Observed state	Digital environment	Each step

RELIABILITY

BASIC KPI INFORMATION		
ID	**Application Domain(s)**	**Name of KPI**
KPI-DF-057	ATM, Power Grid, Railway	Domain shift – success rate drop

Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	17.01.2025	M. Leyli-abadi (IRTSX)	Initial version
1.0	03.03.2025	M. Leyli-abadi (IRTSX)	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	Domain shift – success rate drop KPI measures drop in the performance of the agent after the occurrence of a shift in the source domain.
Objective(s)	To quantify the decline in the agent's performance after a shift in the source domain, providing insights into the agent's ability to maintain effectiveness under altered conditions. This KPI helps in evaluating the agent's resilience, adaptability, and the robustness of its training, facilitating the identification of weaknesses and the development of strategies to improve its performance in dynamic or unpredictable environments. This KPI contributes to evaluating Reliability of the AI-based assistant **when dealing with real-world conditions** which may be slightly different from source domain, as part of Task 4.2 evaluation objectives, and O4 main project objective.
Formula(s)	Schematically, the performance drop is shown in the y-axis of the following plot:  A formula to quantify the drop in performance of the agent after a domain shift could be: $$Performance\ drop = \frac{R_{original} - R_{shifted}}{R_{original}}$$ Where the R could be a performance metric of the AI-based agent like the cumulated Reward. This formula yields a ratio representing the relative drop in performance, with a higher value indicating a more significant drop due to the domain shift.
Unit of Measurement	None

CALCULATION METHODOLOGY		
Baseline calculation methodology		
Step #	**Step description**	**Calculation**

1	Apply all KPI calculation steps and calculate the KPI when source and target domains are distributed following the same distribution (e.g., same topology configuration)	Same as below
KPI calculation methodology		
KPI step #	***Step description***	***Calculation***
1	Generate some observations for source domain	Digital environment
2	Generate some data for target domains from a different distribution than the one used for source domain (the target distribution could be a slightly different configuration of the environment that is not observed or some perturbation of the source domain)	Digital environment and perturbation generation tool
3	Train the RL agent on the source domain	RL training module
4	Evaluate the RL agent on both source and target domains	RL evaluation module
5	Retrieve the cumulated rewards in both domains	Digital environment
6	Compute the performance drop using the formula $Performance\ drop = \frac{R_{original} - R_{shifted}}{R_{original}}$	KPI
7	Report the KPI results. This KPI yields a ratio representing the relative drop in performance, with a higher value indicating a more significant drop due to the domain shift.	Reporting

Data collection			
Data ID	***Type***	***Source***	***Frequency***
$Reward$	Rewards in source and target domains	At the evaluation stage of the RL agent	It could be once, or multiple times if we relate it to the time axis

BASIC KPI INFORMATION			
ID	**Application Domain(s)**	**Name of KPI**	
KPI-DF-054	ATM, Power Grid, Railway	Domain shift – out of domain detection accuracy	
Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	17.01.2025	M. Leyli-abadi (IRTSX)	Initial version
1.0	03.03.2025	M. Leyli-abadi (IRTSX)	Finalization for D4.1

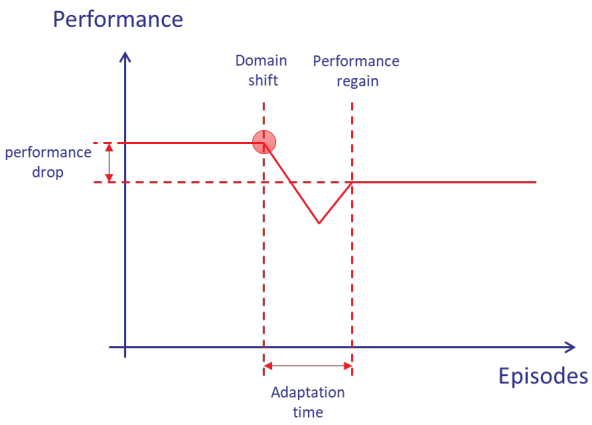
SCOPE AND OBJECTIVES OF KPI	
Description	Domain shift – out of domain detection accuracy measures the accuracy with which the agent can detect whether it is operating in a domain that is different from the one it was trained on. It is useful for systems that need to switch strategies or request human intervention when a domain shift is detected. A recent paper proposed by Nasvytis et al. (2024) introduce various approaches for detection of OOD in RL.
Objective(s)	It is crucial for an AI-based assistant to determine whether it can make reliable decisions in a given configuration. AI algorithms tend to be more dependable when they have been trained on similar configurations. Therefore, if the AI assistant can accurately detect out-of-domain configurations, it can seek human feedback to reduce uncertainty, leading to more adapted and reliable decisions in future scenarios. This KPI allows to determine if AI-based system could detect the shift before decision making. This KPI contributes to evaluating Reliability of the AI-based assistant **when dealing with real-world conditions** which may be slightly different from source domain, as part of Task 4.2 evaluation objectives, and O4 main project objective.
Formula(s)	• If a detection algorithm or tool is used, the accuracy of OOD detection is then given by: $$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$ This formula provides a measure of how well the agent can detect domain shifts, balancing both the correct identification of OOD and ID scenarios. It is essential for systems that need to adapt their strategies or seek human intervention when a domain shift is detected. • Otherwise, compute a distribution-based distance (e.g. Wasserstein) between source and target domains and if this distance is greater than a predefined threshold, we can validate the hypothesis that there is a shift in the data.
Unit of Measurement	Percentage (%) of correctly identified OOD cases

CALCULATION METHODOLOGY		
Baseline calculation methodology		
Step #	**Step description**	**Calculation**
1	Apply all KPI calculation steps and calculate the KPI when source and target domains are distributed following the same distribution (e.g., same topology configuration)	See KPI calculation steps
KPI calculation methodology		
KPI step #	**Step description**	**Calculation**

1	Generate some observations for source and target domains from two different distributions (e.g. source domain data generated using a set of possible network configurations and target domain data generated using a completely different set of configurations, or integrate some perturbations)	Digital environment & perturbation agent
2	Compute the distribution discrepancy between these source and target domains using various existing distribution-based distance measures (Wasserstein distance). Or use an existing tool or algorithm in literature which allows the detection of distribution shift.	Distance measure or a detection algorithm
3	Compute the accuracy of the detector or compare the distance measure with a predefined threshold. • If the detector detects with high accuracy the presence of OOD data, a specific strategy could be adopted by the system to adapt its behavior to the data shift. • Otherwise, if the distance between two distributions is greater than the predefined threshold, we can validate the hypothesis that there is shift in the data and the model should adapt its strategy.	KPI module
4	Report the KPI result. This KPI expresses if there are some distribution shift in the data.	Reporting

Data collection			
Data ID	***Type***	***Source***	***Frequency***
Distance	A real value describing the distance between source and target domains	Before the training of the models, on the raw data	Once, when the data acquisition is finished
Accuracy	A real value describing the precision with which the model was able to detect the OOD data	After the computation of the distance and before the training on the acquired data	Once, when the data acquisition is finished

BASIC KPI INFORMATION			
ID	**Application Domain(s)**	**Name of KPI**	
KPI-DF-052	ATM, Power Grid, Railway	Domain shift - Adaptation time	
Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	17.01.2025	M. Leyli-abadi (IRTSX)	First version
1.0	03.03.2025	M. Leyli-abadi (IRTSX)	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	The time or number of episodes required for the agent to regain a specific level of performance in the shifted domain after the domain shift has occurred. It can be used to evaluate how quickly an agent can adapt to new environmental conditions.
Objective(s)	Domain adaptation (DA) is a sub-field of transfer learning. DA can be defined as the capability to deploy a model trained in one or more source domains into a different target domain. We consider that the source and target domains have the same feature space. In this sense, it is important for RL based agents to have a reasonable adaptation time to a new domain which may present a slight shift from the source domain. However, the adaptation time should also consider the performance drop into its computation, as a high performance drop after the adaptation could not be tolerated. This KPI contributes to evaluating Reliability of the AI-based assistant **when dealing with real-world conditions** which may be slightly different from source domain, as part of Task 4.2 evaluation objectives, and O4 main project objective.
Formula(s)	The adaptation time could be computed as the sum of episodes required for an agent to regain a specific level of performance in the shifted domain after the domain shift has occurred. It could be presented schematically as follows:  In term of formula, we can compute it as: $$\sum_{t_{shift}}^{T} \mathbf{1}\ \{R_{t_{shift}-1} - R_t > \epsilon\}$$ Where **1** is indicator function and becomes one when the difference between the reward computed before the occurrence of the domain shift ($t_{shift} - 1$) and the reward at the time steps after the domain shift ($[t_{shift}; T]$) is higher than a threshold ϵ. The value of the threshold could be dependent to the domain and the sensitivity.
Unit of Measurement	Time, number of time steps, number of episodes

CALCULATION METHODOLOGY		
Baseline calculation methodology		
Step #	**Step description**	**Module(s)**
1	Generate some observations for source and target domains from the same distribution.	Digital environments
2	Train the agent on the source domain	RL module
3	Evaluate the agent on the source and target domains using some test datasets	Evaluation module
4	Compute the performance drop between the source and target domains	KPI
5	If the performance drop is above a predefined threshold, the agent should try to adapt its strategy to target domain	RL module
6	Compute the number of episodes that the agent needs to adapt its behavior to target domain	KPI
7	In baseline, we expect that there is no performance drop and the adaptation time to be zero	Reporting
KPI calculation methodology		
KPI step #	**Step description**	**Module(s)**
1	Generate some observations for source and target domains from two different distributions. (the target distribution could be a slightly different configuration of the environment that is not observed or some perturbation of the source domain)	Digital environment and perturbation generation tool
2	Train the agent on the source domain	RL module
3	Evaluate the agent on the source and target domains using some test datasets	Evaluation module
4	Compute the performance drop between the source and target domains	KPI (Developers in WP2 and WP3)
5	If the performance drop is above a predefined threshold, the agent should try to adapt its strategy to target domain.	RL module
6	Compute the number of episodes that the agent needs to adapt its behavior to target domain by computing the performance drop at each episode	KPI (Developers in WP2 and WP3)
7	Report the number of episodes if the agent were able to adapt its behavior, otherwise report an analysis on the behavior of the agent (performance drop and time passed for adaptation without success).	Reporting

Data collection			
Data ID	**Type**	**Source**	**Frequency**
$AdaptationTime$	Time required by a RL agent for adaptation to a shifted domain	In the RL environment At the evaluation phase of the agent on the shift domain	As long as the agent may readapt its strategy to attain a satisfactory performance
$PerformanceDrop$	Performance drop of RL agent	In the RL environment At the evaluation phase of the agent on the shift domain	As long as the agent may readapt its strategy to attain a satisfactory performance

BASIC KPI INFORMATION		
ID	**Application Domain(s)**	**Name of KPI**
KPI-DF-090	ATM, Power Grid, Railway	Domain shift – Forgetting rate

Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	03.03.2025	M. Leyli-abadi (IRTSX)	First version
1.0	04.03.2025	M. Leyli-abadi (IRTSX)	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	The rate at which an agent forgets its performance in the original domain after being exposed to a shifted domain. It helps to measure the extent to which learning in the new domain negatively impacts the agent's ability to perform in the original domain.
Objective(s)	The objective of computing the *Forgetting Rate in Domain Shift* is to quantify the decline in an agent's performance on the original domain after being trained or exposed to a shifted domain. This metric helps assess the extent of negative transfer, ensuring that adaptation to the new domain does not excessively degrade prior knowledge. A higher forgetting rate indicates a more significant loss of previously learned knowledge due to domain shift. This KPI contributes to evaluating Reliability of the AI-based assistant **when dealing with real-world conditions** which may be slightly different from source domain, as part of Task 4.2 evaluation objectives, and O4 main project objective.
Formula(s)	Let: • P_{orig}^{init} be the agent's performance (e.g., accuracy, reward, or another metric) in the original domain before exposure to the new domain. • P_{orig}^{post} be the agent's performance in the original domain after training in the shifted domain. The forgetting rate (FR) can be computed as: $$FR = \frac{P_{orig}^{init} - P_{orig}^{post}}{P_{orig}^{init}}$$ **Interpretation** • $FR = 0 \rightarrow$ No forgetting; the model retains full performance in the original domain. • $0 < FR \leq 1 \rightarrow$ some forgetting; a drop in performance • $FR > 1 \rightarrow$ sever forgetting; performance drops below zero (e.g., when performance is normalized and turns negative).
Unit of Measurement	Proportion or Percentage.

CALCULATION METHODOLOGY		
Baseline calculation methodology		
Step #	**Step description**	**Module(s)**
1	Apply all KPI calculation steps and calculate the KPI when source and target domains are distributed following the same distribution (e.g., same topology configuration)	Same as below
KPI calculation methodology		
KPI step #	**Step description**	**Module(s)**
1	Generate some observations for source and target domains from two different distributions. (the target distribution could be a slightly different configuration of the environment that is not observed or some perturbation applied on the source domain)	Digital environment and perturbation generation tool

2	Train the agent on the source domain	RL module
3	Evaluate the agent on the source domain and compute the related performance (e.g., reward) and call it P_{orig}^{init}	Evaluation module
4	Train the same agent on the shifted domain dataset for adaptation	RL module
5	Evaluate its performance on another test dataset driven from the source domain and call it P_{orig}^{post}	Evaluation module
6	Compute the forgetting rate KPI as $$FR = \frac{P_{orig}^{init} - P_{orig}^{post}}{P_{orig}^{init}}$$	KPI (Developers in WP2 and WP3)
7	Report the KPI using the interpretation given in the formula section above.	Reporting

Data collection			
Data ID	***Type***	***Source***	***Frequency***
$Performance$	Reward obtained by an RL agent on two different datasets driven from the source domain	In the RL environment at the evaluation phase of the agent on source domain	Twice. Once the agent is trained only on the source domain, and one additional time when the agent is retrained on the target domain.

BASIC KPI INFORMATION		
ID	**Application Domain(s)**	**Name of KPI**
KPI-DF-056	ATM, Power Grid, Railway	Domain shift – robustness to domain parameters

Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	17.01.2025	M. Leyli-abadi (IRTSX)	Initial version
1.0	03.03.2025	M. Leyli-abadi (IRTSX)	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	Robustness to domain parameters KPI evaluates the sensitivity of the agent's performance (e.g., Reward) to changes in specific domain parameters (e.g., generators type including renewables in power grid domain). It helps to identify which environmental factors most affect the agent's performance.
Objective(s)	To assess the sensitivity of the agent's performance to variations in domain parameters, identifying key environmental factors that significantly impact the agent's effectiveness and robustness, thereby guiding improvements in adaptability and resilience across different scenarios. This KPI contributes to evaluating Reliability of the AI-based assistant **when dealing with real-world conditions** which may be slightly different from source domain, as part of Task 4.2 evaluation objectives, and O4 main project objective.
Formula(s)	Calculating the variance or standard deviation of the rewards obtained by the agent after introducing changes in the source domain and comparing it to the standard deviation before the changes, can provide insights into the robustness of the agent's performance under varying domain parameters. To formalize the definition, let: • R_{before} represent the rewards obtained by the agent before introducing changes. • R_{after} represent the rewards obtained after introducing changes. • σ_{before} be the standard deviation of R_{before}. • σ_{after} be the standard deviation of R_{after}. • $\Delta\sigma$ be the difference between the two standard deviations. The formula to quantify the change in variability due to domain changes is: $$\Delta\sigma = \sigma_{after} - \sigma_{before}$$ Where each standard deviation σ is calculated as: $$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(R_i - \bar{R})^2}$$ with R_i being individual rewards and $\bar{R}$ the mean reward for the respective domain state (before or after changes). The difference $\Delta\sigma$ gives an indication of the robustness of the agent's performance in response to changes in the domain parameters.
Unit of Measurement	None

CALCULATION METHODOLOGY		
Baseline calculation methodology		
Step #	**Step description**	**Calculation**

1	Apply all KPI calculation steps and calculate the KPI when source and target domains are distributed following the same distribution (e.g., same topology configuration)	Same as below
KPI calculation methodology		
KPI step #	***Step description***	***Calculation***
1	Generate some data from source domain (a set of basic configurations)	Digital environment
2	Generate data for target domain by changing one of domain parameters (e.g., number of renewables, power line disconnections, etc.)	Digital environment
3	Train the agent on the source domain	RL training module
4	Evaluate the agent on the source and target domains to retrieve the corresponding rewards	RL evaluation module
5	Compute the standard deviation of each individual reward on a specific state (observation) with respect to the mean reward and that for each of the source and target domains separately $\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(R_i - \bar{R})^2}$	KPI
6	Compute the difference between the standard deviations computed the previous step $\Delta\sigma = \sigma_{after} - \sigma_{before}$	KPI
7	Report the result as the variation (robustness) of the agent's performance with respect to the changes in domain parameters.	Reporting

Data collection			
Data ID	***Type***	***Source***	***Frequency***
$Reward$	Individual and mean rewards of the agents in both source and target domains	At the evaluation stage of the RL agent	Once, after the training of the agent and at the evaluation stage on both source and target domains

BASIC KPI INFORMATION			
ID	**Application Domain(s)**	**Name of KPI**	
KPI-DF-055	ATM, Power Grid, Railway	Domain shift – Policy robustness	
Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	17.01.2025	M. Leyli-abadi (IRTSX)	Initial version
1.0	03.03.2025	M. Leyli-abadi (IRTSX)	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	Domain shift – Policy robustness KPI calculates a ratio of the performance in the shifted domain to the performance in the original domain. A score close to 1 indicates high robustness, while a lower score indicates reduced performance due to the domain shift. It can be used to assess the generalization of a policy learned in a simulated environment when applied to a real-world scenario.
Objective(s)	To evaluate the robustness and generalization capability of a policy by measuring its performance ratio between a shifted domain and the original domain, ensuring that policies trained in simulated environments maintain high effectiveness when applied to real-world scenarios. This KPI contributes to evaluating Reliability of the AI-based assistant **when dealing with real-world conditions** which may be slightly different from source domain, as part of Task 4.2 evaluation objectives, and O4 main project objective.
Formula(s)	If we present by R_{shift} the performance or reward obtained in shifted domain and by $R_{original}$ the performance or reward in the source domain, the ratio is computed as: $$Policy\ Robustness = \frac{R_{shifted}}{R_{original}}$$
Unit of Measurement	None

CALCULATION METHODOLOGY		
Baseline calculation methodology		
Step #	**Step description**	**Calculation**
1	Apply all KPI calculation steps and calculate the KPI when source and target domains are distributed following the same distribution (e.g., same topology configuration)	Same as below
KPI calculation methodology		
KPI step #	**Step description**	**Calculation**
1	Generate some observations for source domain. Consider some real-world situations or data as target domain which represents the same set of features	Digital environment
2	Train the agent on the source domain.	RL module
3	Evaluate the agent's performance on the source distribution and get the cumulated reward ($R_{original}$)	Evaluation module
4	Evaluate the agent's performance on the target domain with real-world data and get the associated cumulated reward ($R_{shifted}$)	Evaluation module
5	Compute the policy robustness on the rewards computed in step 3 and 4	KPI

6	Report the result as an indicator of agent's robustness to real-world situation (shifted target). A score close to 1 indicates the high robustness.	Reporting

Data collection			
Data ID	**Type**	**Source**	**Frequency**
$Reward$	Rewards of the agents in two different situations	At the evaluation stage of an RL agent	Once, after the training of the agent and the evaluation stage.

BASIC KPI INFORMATION			
ID	**Application Domain(s)**	**Name of KPI**	
KPI-DF-053	ATM, Power Grid, Railway	Domain shift – generalization gap	
Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	17.01.2025	M. Leyli-abadi (IRTSX)	First version
1.0	03.03.2025	M. Leyli-abadi (IRTSX)	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	Domain shift – generalization gap evaluates the absolute difference between the performance (e.g., rewards) in the training domain and the shifted domain. This metrics quantifies the extent of performance loss due to domain shift.
Objective(s)	The objective is to verify to which extent the AI-based assistant performance deteriorates when the target domain presents some changes in comparison to the source domain. If an agent can retain the same performance expectations in shifted domain, it will be qualified as reliable. This KPI contributes to evaluating Reliability of the AI-based assistant **when dealing with real-world conditions** which may be slightly different from source domain, as part of Task 4.2 evaluation objectives, and O4 main project objective.
Formula(s)	$Generalization\ Gap = \left\lvert R_{source\ domain} - R_{target\ domain} \right\rvert$
Unit of Measurement	No units

CALCULATION METHODOLOGY		
Baseline calculation methodology		
Step #	**Step description**	**Calculation**
	Apply all KPI calculation steps and calculate the KPI when source and target domains are distributed following the same distribution (e.g., same topology configuration)	Same as below
KPI calculation methodology		
KPI step #	**Step description**	**Calculation**
1	Generate some observations for source and target domains from two different distributions. (the target distribution could be a slightly different configuration of the environment that is not observed or some perturbation of the source domain)	Digital environment and perturbation generation tool
2	Train an agent on the source domain	RL module
3	Evaluate the agent on the source (nominal behavior) domain and get the reward	Evaluation module
4	Evaluate the agent on the target (perturbed or out-of-domain) domain and get the reward	Evaluation module
5	Compute the absolute difference between these two rewards	KPI
6	Compare the KPI value with the baseline and report the metric as the generalization gap	Report (Developers in WP2 and WP3)

Data collection			
Data ID	**Type**	**Source**	**Frequency**
$GeneralizationGap$	A real value describing the generalization gap in domain shift	During the training and evaluation of the agent	Once the agent is trained, it could be computed once for two domains

SOCIAL-TECHNICAL DECISION QUALITY

This annex details all evaluation protocols corresponding to §5 - Social-technical decision quality, and Social-technical decision quality, AI acceptability, trust, and trustworthiness, Human user experience, AI-human learning curve, AI-human task allocation balance, Long-term consequences of AI assistants objectives.

Page intentionally left blank.

SOCIAL-TECHNICAL DECISION QUALITY

BASIC KPI INFORMATION		
ID	**Application Domain(s)**	**Name of KPI**
KPI-HS-003	ATM, Power Grid, Railway	Human Intervention Frequency

Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	07.02.2025	Roman Liessner (DB)	Creation of the document
1.0	03.03.2025	Roman Liessner (DB)	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	The **Human Intervention Frequency** KPI measures the proportion of instances in which a human operator intervenes in an automated decision-making process. While this KPI was initially developed for railway traffic control scenarios, it has been generalized to assess the reliability and autonomy of any AI-assisted system. It reflects the trust placed in the AI by quantifying how often human corrections are required during routine operations.
Objective(s)	This KPI contributes to evaluating Social-technical decision quality of the AI-based assistant, as part of Task 4.3 evaluation objectives, and O3 main project objective: • To evaluate the effectiveness of the AI system in operating autonomously. • To provide a performance benchmark for minimizing human interventions across various domains. • To identify areas where the AI may require additional refinement or support.
Formula(s)	$AcceptanceScore = \frac{n_{decision}^{operator}}{n_{decision}^{AI}} \times 100$
Unit of Measurement	**Percentage (%)** of AI decisions requiring human intervention.

CALCULATION METHODOLOGY		
Baseline calculation methodology		
Step #	**Step description**	**Calculation**
1	Define a baseline by monitoring human interventions in current railway operations without AI support. .	Network Operator
2	Establish historical data of human decision-making in re-scheduling events.	Data Analysts
KPI calculation methodology		
KPI step #	**Step description**	**Calculation**
1	Monitor AI-based scheduling in real-time and log all decision instances.	AI System
2	Log instances where human operators override or adjust AI decisions.	Railway Control Center
3	Compute the acceptance score using the formula above.	Data Analysts
4	Compare acceptance scores over time to track improvements in AI reliability.	Network Operator

Data collection			
Data ID	**Type**	**Source**	**Frequency**
$n_{decision}^{AI}$	Total number of AI decision instances	Recommendation module	Each step (realtime)

$n_{decision}^{operator}$	Number of AI decision instances where human operators override or adjust AI decisions.	Recommendation module	Each step (realtime)

BASIC KPI INFORMATION			
ID	**Application Domain(s)**	**Name of KPI**	
KPI-SS-030	ATM, Power Grid, Railway	Significance of human revisions	
Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	07.03.2025	Kostiantyn Kucher (LiU)	Initial version
1.0	08.03.2025	Kostiantyn Kucher (LiU)	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	This KPI represents human operators' subjective assessment of necessary revisions for the AI-generated solutions by the human operator, self-reported by the operator with Likert-scale questions.
Objective(s)	This KPI contributes to evaluating Social-technical decision quality of the AI-based assistant, as part of Task 4.3 evaluation objectives, and O3 main project objective. It is also relevant to protocols and concepts defined in D1.1 such as "Significance of human revisions".
Formula(s)	As operationalized by the questionnaire (normally Likert-scales with several items that are rated on a scale of e.g. 1–5 or 1–7).
Unit of Measurement	Ordinal data response on a Likert scale (or potentially a similar response on an interval scale)

CALCULATION METHODOLOGY		
Baseline calculation methodology		
Step #	**Step description**	**Calculation**
1	An existing questionnaire such as ATAI (Sindermann, et al., 2021) can be used to capture the self-reported pre-study attitude towards the AI assistant as an indirect baseline.	Simulation and Testing Tools (Operations Testing Team)
KPI calculation methodology		
KPI step #	**Step description**	**Calculation**
1	The KPI is measured by the means of a questionnaire comprising one or several questions on a Likert scale. The basic version of the questionnaire could consist of one or several questions / statements to be rated such as "The solution/decision proposed by the AI assistant required the following level of revisions: no / minor / medium / major" and "The solution/decision proposed by the AI assistant required the following number of revisions: no / low / medium / high". Further questions could be adapted from the XAI for Human-Agent Interaction Survey (Silva, et al., 2022) as well as the work on contradictory decisions and explanations (Ebermann, et al., 2023). The contents of the questionnaire should be adjusted and initially tested according to the particular experimental design.	Simulation and Testing Tools (Operations Testing Team)

Data collection			
Data ID	**Type**	**Source**	**Frequency**
questionnaire	Questionnaire with several questions on a Likert scale	A novel questionnaire designed specifically for this study, or adapted from the existing work, such as the XAI for Human-Agent Interaction Survey (Silva, et al., 2022) as well as the work on contradictory decisions and explanations (Ebermann, et al., 2023).	In dependency of the particular experimental design

BASIC KPI INFORMATION			
ID	**Application Domain(s)**	**Name of KPI**	
KPI-PS-089	ATM, Power Grid, Railway	Perceived decision novelty	
Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	07.03.2025	Kostiantyn Kucher (LiU)	Initial version
1.0	08.03.2025	Kostiantyn Kucher (LiU)	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	This KPI represents human operators' self-reported subjective assessment of nontriviality for the AI-generated solutions measured with a questionnaire.
Objective(s)	This KPI contributes to evaluating Social-technical decision quality of the AI-based assistant, as part of Task 4.3 evaluation objectives, and O3 main project objective.
Formula(s)	As operationalized by the questionnaire (normally Likert-scales with several items that are rated on a scale of e.g. 1–5 or 1–7).
Unit of Measurement	Ordinal data response on a Likert scale (or potentially a similar response on an interval scale)

CALCULATION METHODOLOGY		
Baseline calculation methodology		
Step #	**Step description**	**Calculation**
1	An existing questionnaire such as ATAI (Sindermann et al., 2021) can be used to capture the self-reported pre-study attitude towards the AI assistant as an indirect baseline.	Simulation and Testing Tools (Operations Testing Team)
KPI calculation methodology		
KPI step #	**Step description**	**Calculation**
1	The KPI is measured by the means of a questionnaire comprising one or several questions on a Likert scale. The basic version of the questionnaire could consist of one or several questions / statements to be rated such as "I find the solution/decision proposed by the AI assistant to be non-trivial" and "I find that the solution/decision proposed by the AI assistant creatively complements my own solution/decision". Further questions could be adapted from the Human-Computer Trust questionnaire (Madsen, et al., 2000) or the Explanation Satisfaction Scale (Hoffman, et al., 2023). The contents of the questionnaire should be adjusted and initially tested according to the particular experimental design.	Simulation and Testing Tools (Operations Testing Team)

Data collection			
Data ID	**Type**	**Source**	**Frequency**
questionnaire	Questionnaire with one or several questions on a Likert scale	A novel questionnaire designed specifically for this study or adapted from the existing work, such as the Human-Computer Trust questionnaire (Madsen, et al., 2000) or the Explanation Satisfaction Scale (Hoffman, et al., 2023).	In dependency of the particular experimental design

AI ACCEPTABILITY, TRUST, AND TRUSTWORTHINESS

BASIC KPI INFORMATION		
ID	**Application Domain(s)**	**Name of KPI**
KPI-AS-002	ATM, PowerGrid, Railway	Acceptance

Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	03.02.2025	Patrick Zinsli (FHNW)	Document creation
1.0	03.03.2025	Patrick Zinsli (FHNW)	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	Acceptance of the system by a human user.
Objective(s)	This KPI contributes to evaluating AI acceptability, trust and trustworthiness of the AI-based assistant, as part of Task 4.3 evaluation objectives, and O2 main project objective.
Formula(s)	Using a TAM model (technology acceptance model) or similar e.g. the AI-Acceptance model (KIAM) (Scheuer, 2020)
Unit of Measurement	As operationalized by the questionnaire (normally Likert-scales with several items that are rated on a scale of e.g. 1-5)

CALCULATION METHODOLOGY		
Baseline calculation methodology		
Step #	**Step description**	**Calculation**
1	Baseline is measured in an experimental setting with test persons in the control group (to see whether it differs from the experimental group). This requires a sufficient sample size.	Simulation and Testing Tools (operations Testing Team)
KPI calculation methodology		
KPI step #	**Step description**	**Calculation**
1	In dependency of the particular experimental design, the KPI is measured once or several times at defined points of time with test persons of the control group as well as of the experimental group by the means of a standardized questionnaire.	Simulation and Testing Tools (operations Testing Team)

Data collection			
Data ID	**Type**	**Source**	**Frequency**
questionnaire	Standardized questionnaire	The Acceptance of the system by a human user can be measured using the AI-Acceptance model (KIAM) (Scheuer, 2020) or similar.	In dependency of the particular experimental design

BASIC KPI INFORMATION		
ID	**Application Domain(s)**	**Name of KPI**
KPI-TS-039	ATM, Power Grid, Railway	Trust towards the AI tool

Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	29.01.2025	Patrick Zinsli (FHNW)	Document creation
1.0	03.03.2025	Patrick Zinsli (FHNW)	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	(Dis)trust is defined here as a sentiment resulting from knowledge, beliefs, emotions, and other elements derived from lived or transmitted experience, which generates positive or negative expectations concerning the reactions of a system and the interaction with it (whether it is a question of another human being, an organization or a technology)" ((Cahour, et al., 2009), p. 1261).
Objective(s)	This KPI contributes to evaluating AI acceptability, trust and trustworthiness of the AI-based assistant, as part of Task 4.3 evaluation objectives, and O2 main project objective.
Formula(s)	As operationalized by the questionnaire (normally Likert-scales with several items that are rated on a scale of e.g. 1-5)
Unit of Measurement	Lickert-Scale or similar

CALCULATION METHODOLOGY		
Baseline calculation methodology		
Step #	**Step description**	**Calculation**
1	Baseline is measured in an experimental setting with test persons in the control group (to see whether it differs from the experimental group). This requires a sufficient sample size.	Simulation and Testing Tools (Operations Testing Team)
KPI calculation methodology		
KPI step #	**Step description**	**Calculation**
1	In dependency of the particular experimental design, the KPI is measured once or several times at defined points of time with test persons of the control group as well as of the experimental group by the means of a standardized questionnaire.	Simulation and Testing Tools (Operations Testing Team)

Data collection			
Data ID	**Type**	**Source**	**Frequency**
questionnaire	Standardized questionnaire	The human operators' trust towards the AI tool can be measured using the Scale for XAI (Hoffman, et al., 2018) or similar.	In dependency of the particular experimental design

BASIC KPI INFORMATION			
ID	**Application Domain(s)**	**Name of KPI**	
KPI-AS-005	ATM, Power Grid, Railway	Agreement score	
Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	07.03.2025	Kostiantyn Kucher (LiU)	Initial version
1.0	08.03.2025	Kostiantyn Kucher (LiU)	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	This KPI represents human operators' self-reported agreement with individual AI-generated solutions/decisions on a scale of 0–100.
Objective(s)	This KPI contributes to evaluating AI acceptability, trust and trustworthiness of the AI-based assistant, as part of Task 4.3 evaluation objectives, and O2 main project objective. It is also relevant to protocols and concepts defined in D1.1 such as "Agreement score".
Formula(s)	Self-reported agreement with specific solutions on a scale of 0–100.
Unit of Measurement	Interval scale response

CALCULATION METHODOLOGY			
Baseline calculation methodology			
Step #	**Step description**		**Calculation**
1	An existing questionnaire such as ATAI (Sindermann, et al., 2021) can be used to capture the self-reported pre-study attitude towards the AI assistant as an indirect baseline.		Simulation and Testing Tools (Operations Testing Team)
KPI calculation methodology			
KPI step #	**Step description**		**Calculation**
1	The KPI is measured by capturing an explicit numerical agreement score from the user for a given solution/decision.		Simulation and Testing Tools (Operations Testing Team)
Data collection			
Data ID	**Type**	**Source**	**Frequency**
score	Numerical agreement score on a scale 0–100	Explicit input from the user.	In dependency of the particular experimental design

BASIC KPI INFORMATION			
ID	**Application Domain(s)**	**Name of KPI**	
KPI-TS-038	ATM, Power Grid, Railway	Trust in AI solutions score	
Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	07.03.2025	Kostiantyn Kucher (LiU)	Initial version
1.0	08.03.2025	Kostiantyn Kucher (LiU)	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	This KPI represents human operators' self-reported trust (attitude) for individual AI-generated solutions measured with a questionnaire.
Objective(s)	This KPI contributes to evaluating AI acceptability, trust and trustworthiness of the AI-based assistant, as part of Task 4.3 evaluation objectives, and O2 main project objective. It is also relevant to protocols and concepts defined in D1.1 such as "Trust in AI solutions score".
Formula(s)	As operationalized by the questionnaire (normally Likert-scales with several items that are rated on a scale of e.g. 1–5 or 1–7).
Unit of Measurement	Ordinal data response on a Likert scale (or potentially a similar response on an interval scale)

CALCULATION METHODOLOGY		
Baseline calculation methodology		
Step #	**Step description**	**Calculation**
1	An existing questionnaire such as ATAI (Sindermann et al., 2021) can be used to capture the self-reported pre-study attitude towards the AI assistant as an indirect baseline.	Simulation and Testing Tools (Operations Testing Team)
KPI calculation methodology		
KPI step #	**Step description**	**Calculation**
1	The KPI is measured by the means of a questionnaire comprising one or several questions on a Likert scale. The basic version of the questionnaire could consist of several questions / statements to be rated such as "I trust the solution/decision proposed by the AI assistant","I find the solution/decision proposed by the AI assistant to be coherent with", and "I find the solution/decision proposed by the AI assistant to be trustworthy without further additional explanations". Further questions could be adapted from the Explanation Satisfaction Scale (Hoffman, et al., 2023), the XAI for Human-Agent Interaction Survey (Silva, et al., 2022) as well as Co-12 Explanation Properties (Nauta, et al., 2023). The contents of the questionnaire should be adjusted and initially tested according to the particular experimental design.	Simulation and Testing Tools (Operations Testing Team)

Data collection			
Data ID	**Type**	**Source**	**Frequency**
questionnaire	Questionnaire with several questions on a Likert scale	A novel questionnaire designed specifically for this study or adapted from the existing work, such as the Explanation Satisfaction Scale (Hoffman, et al., 2023), the XAI for Human-Agent Interaction Survey (Silva, et al., 2022), as well as Co-12 Explanation Properties (Nauta, et al., 2023).	In dependency of the particular experimental design

BASIC KPI INFORMATION		
ID	**Application Domain(s)**	**Name of KPI**
KPI-CS-013	ATM, Power Grid, Railway	Comprehensibility

Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	07.03.2025	Kostiantyn Kucher (LiU)	Initial version
1.0	08.03.2025	Kostiantyn Kucher (LiU)	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	This KPI represents human operators' self-reported ability to understand and thus make use of the AI-generated solution/decision, measured with a questionnaire.
Objective(s)	This KPI contributes to evaluating AI acceptability, trust and trustworthiness of the AI-based assistant, as part of Task 4.3 evaluation objectives, and O2 main project objective. It is also relevant to protocols and concepts defined in D1.1 such as "Comprehensibility".
Formula(s)	As operationalized by the questionnaire (normally Likert-scales with several items that are rated on a scale of e.g. 1–5 or 1–7).
Unit of Measurement	Ordinal data response on a Likert scale (or potentially a similar response on an interval scale)

CALCULATION METHODOLOGY		
Baseline calculation methodology		
Step #	**Step description**	**Calculation**
1	An existing questionnaire such as ATAI (Sindermann et al., 2021) can be used to capture the self-reported pre-study attitude towards the AI assistant as an indirect baseline.	Simulation and Testing Tools (Operations Testing Team)
KPI calculation methodology		
KPI step #	**Step description**	**Calculation**
1	The KPI is measured by the means of a questionnaire comprising one or several questions on a Likert scale. The basic version of the questionnaire could consist of one or several questions / statements to be rated such as "I understand the logic behind the solution/decision proposed by the AI assistant". Further questions could be adapted from the Explanation Satisfaction Scale (Hoffman, et al., 2023), the XAI for Human-Agent Interaction Survey (Silva, et al., 2022) as well as Co-12 Explanation Properties (Nauta, et al., 2023). The contents of the questionnaire should be adjusted and initially tested according to the particular experimental design.	Simulation and Testing Tools (Operations Testing Team)

Data collection			
Data ID	**Type**	**Source**	**Frequency**
questionnaire	Questionnaire with several questions on a Likert scale	A novel questionnaire designed specifically for this study or adapted from the existing work, such as the Explanation Satisfaction Scale (Hoffman, et al., 2023), the XAI for Human-Agent Interaction Survey (Silva, et al., 2022), as well as Co-12 Explanation Properties (Nauta, et al., 2023).	In dependency of the particular experimental design

HUMAN USER EXPERIENCE

BASIC KPI INFORMATION		
ID	**Application Domain(s)**	**Name of KPI**
KPI-WS-040	ATM, Power Grid, Railway	Workload

Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	31.10.2024	Duarte Dias (INESC TEC)	Creation of the document
1.0	03.03.2025	Duarte Dias (INESC TEC)	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	Workload KPI is based on the workload assessment of human operators of the AI assistant. After each testing session using the system, the workload of human operators due to the AI assistant will be evaluated to understand in which scenarios (and depending on the AI level of support) it contributes for a higher workload.
Objective(s)	This KPI assesses whether the inputs of the operators are according to their real psychophysiology. This can act as a verification methodology but also support the AI to adapt. This KPI will be analyzed together with the "Impact on workload" KPI-IS-041. This KPI contributes to evaluating Human-user experience of the AI-based assistant, as part of Task 4.3 evaluation objectives, and O3 main project objective.
Formula(s)	It shall be determined according to the NASA-TLX methodology or similar. This KPI is still under analysis on how it will be implemented. If with a single manual questionnaire or with a pop-up in the dashboard.
Unit of Measurement	None (rating scale)

CALCULATION METHODOLOGY		
Baseline calculation methodology		
Step #	**Step description**	**Calculation**
1	Baseline data collection procedure – data collected outside the working environment, during a resting period .	See KPI calculation methodology
2	Standardized experimental stress procedure – Trier Social Stress Test - TSST, along with a 2-choice Reaction Time Task - that was successfully implemented by the research team in similar settings	See KPI calculation methodology
KPI calculation methodology		
KPI step #	**Step description**	**Calculation**
1	Operator monitoring in real-time during the different tests with collection of physiological data and computation of biomarkers in real-time to provide to the AI decision support system.	To be defined

Data collection			
Data ID	**Type**	**Source**	**Frequency**
n.a.	n.a.	n.a.	n.a.

BASIC KPI INFORMATION		
ID	**Application Domain(s)**	**Name of KPI**
KPI-AS-009	ATM, Power Grid, Railway	Assistant disturbance

Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	31.10.2024	Duarte Dias (INESC TEC)	Creation of the document
1.0	03.03.2025	Duarte Dias (INESC TEC)	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	Assistant disturbance KPI aims to measure if the AI assistant's notifications are disturbing the human operator's activity.
Objective(s)	This KPI assesses whether the inputs of the operators are according to their real psychophysiology. This can act as a verification methodology but also support the AI to adapt. This KPI contributes to evaluating Human-user experience of the AI-based assistant, as part of Task 4.3 evaluation objectives, and O3 main project objective.
Formula(s)	For each notification, the score ranges in [0, 5] with: • 0 meaning that the notification was not considered disturbing at all by the human operator • 5 means that the human operator considered the notification as fully disturbing This KPI is still under analysis on how it will be implemented. If with a single manual questionnaire or with a pop-up in the dashboard.
Unit of Measurement	None (score)

CALCULATION METHODOLOGY		
Baseline calculation methodology		
Step #	**Step description**	**Calculation**
1	Baseline data collection procedure – data collected outside the working environment, during a resting period .	See KPI calculation methodology
2	Standardized experimental stress procedure – Trier Social Stress Test - TSST, along with a 2-choice Reaction Time Task - that was successfully implemented by the research team in similar settings	See KPI calculation methodology
KPI calculation methodology		
KPI step #	**Step description**	**Calculation**
1	Operator monitoring in real-time during the different tests with collection of physiological data and computation of biomarkers in real-time to provide to the AI decision support system.	To be defined

Data collection			
Data ID	**Type**	**Source**	**Frequency**
n.a.	n.a.	n.a.	n.a.

BASIC KPI INFORMATION			
ID	**Application Domain(s)**	**Name of KPI**	
KPI-CS-049	ATM, PowerGrid, Railway	Cognitive Performance & Stress	
Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	23.01.2025	Duarte Dias (INESC TEC)	First full description of the KPI
1.0	03.03.2025	Duarte Dias (INESC TEC)	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	Cognitive Performance & Stress KPI performs an implicit assessment of the human cognitive performance status and stress levels along the different task that will be performed. The output provides information about the operator mental status and aims to be used to integrate the AI system to contribute as a reward to better adapt decision system.
Objective(s)	The computation of the metrics will be made on the Human Assessment Module and will be integrated in the system that will Tune the autonomy Level of the system. Taking this into account, the objective is to be able to tune the system autonomy level based on the implicit assessment in real time. For example, higher traffic or hard situations/decisions will be detected with any interference with the human operator, implicitly providing information to be used by the decision system. This KPI will not focus on the final results when this module is integrated, but in the calculation of personalized cognitive and stress metrics of a single human based on an individual assessment protocol. If we are not able to perform such protocol, then this module will be generic and not personalized, removing this KPIs. In the personalization we aim to achieve a 20-30% improvement on performance of the model based for a single individual data, enabling a high level of personalization. This KPI contributes to evaluating Human-user experience of the AI-based assistant, as part of Task 4.3 evaluation objectives, and O3 main project objective.
Formula(s)	Performance of the model to predict cognitive status and stress of a single user.
Unit of Measurement	Percentage (%)

CALCULATION METHODOLOGY		
Baseline calculation methodology		
Step #	**Step description**	**Calculation**
1	An experimental protocol was designed to be able to train the model and personalize it to a single individual. This protocol is based on a TSST test combined with a reaction time test. The implementation of such protocol will allow to refine the Human Assessment Module for that specific individual with a direct impact in the Tuning of the autonomy level of the AI system with performing the tasks.	Human Assessment Module (Duarte Dias, INESC TEC)
KPI calculation methodology		
KPI step #	**Step description**	**Calculation**
1	The data collected in the experimental protocol will be used in the Human Assessment Module before and after the personalization. The performance increment will be calculated in %.	Human Assessment Module (Duarte Dias, INESC TEC)

Data collection			
Data ID	**Type**	**Source**	**Frequency**
$DS.2.1$	Experimental and operational data collection for personalized biomarkers	Personal data from a Dataset from INESC TEC	Data previously collected
New data to be collected 1	Experimental data collection for personalized module	Wearable device from INESC TEC jointly with analog visual scales and rection time software will be used to collect such data.	Will depend on the operator's acceptance to participate in the study
New data to be collected 2	Operation data collection for real time tune of the system autonomy	Wearable device from INESC TEC will be used to collect such data.	Will depend on the operator's acceptance to participate in the study

BASIC KPI INFORMATION		
ID	**Application Domain(s)**	**Name of KPI**
KPI-AS-001	ATM, PowerGrid, Railway	Ability to anticipate

Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	06.02.2025	Patrick Zinsli (FHNW)	Document creation
1.0	03.03.2025	Patrick Zinsli (FHNW)	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	"*The ability to anticipate. Knowing what to expect or being able to anticipate developments further into the future, such as potential disruptions, novel demands or constraints, new opportunities, or changing operating conditions*" (Hollnagel, 2015), p. 4 The human operator's ability to anticipate further into the future can be measured by calculating the ratio of (proactively) prevented deviations to actual deviations. In addition, the extent to which the anticipatory sensemaking process of the human operator is supported by AI-based assistants can be measured using the Rigor-Metric for Sensemaking (Zelik, et al., 2018) or similar. The instrument needs to be further developed and adapted to the AI context.
Objective(s)	This KPI contributes to evaluating Human user experience of the AI-based assistant, as part of Task 4.3 evaluation objectives, and O3 main project objective.
Formula(s)	As operationalized by the questionnaire (normally Likert-scales with several items that are rated on a scale of e.g. 1-5)
Unit of Measurement	Lickert-Scale or similar

CALCULATION METHODOLOGY		
Baseline calculation methodology		
Step #	**Step description**	**Calculation**
1	Baseline is measured in an experimental setting with test persons in the control group (to see whether it differs from the experimental group). This requires a sufficient sample size.	Simulation and Testing Tools (operations Testing Team)
KPI calculation methodology		
KPI step #	**Step description**	**Calculation**
1	In dependency of the particular experimental design, the KPI is measured once or several times at defined points of time with test persons of the control group as well as of the experimental group by the means of a standardized questionnaire.	Simulation and Testing Tools (operations Testing Team)

Data collection			
Data ID	**Type**	**Source**	**Frequency**
questionnaire	Standardized questionnaire	Rigor-Metric for Sensemaking (Zelik, et al., 2018) or similar	In dependency of the particular experimental design

BASIC KPI INFORMATION		
ID	**Application Domain(s)**	**Name of KPI**
KPI-SS-031	ATM, Power Grid, Railway	Situation awareness

Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	06.02.2025	Patrick Zinsli (FHNW)	Creating document
1.0	03.03.2025	Patrick Zinsli (FHNW)	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	"Situation Awareness is the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future" (Endsley, 1988).
Objective(s)	This KPI contributes to evaluating Human-user experience of the AI-based assistant, as part of Task 4.3 evaluation objectives, and O3 main project objective.
Formula(s)	As operationalized by the questionnaire (normally Likert-scales with several items that are rated on a scale of e.g. 1-5)
Unit of Measurement	Lickert-Scale or similar

CALCULATION METHODOLOGY		
Baseline calculation methodology		
Step #	**Step description**	**Calculation**
1	Baseline is measured in an experimental setting with test persons in the control group (to see whether it differs from the experimental group). This requires a sufficient sample size.	Simulation and Testing Tools (Operations Testing Team)
KPI calculation methodology		
KPI step #	**Step description**	**Calculation**
1	In dependency of the particular experimental design, the KPI is measured once or several times at defined points of time with test persons of the control group as well as of the experimental group by the means of a standardized questionnaire.	Simulation and Testing Tools (Operations Testing Team)

Data collection			
Data ID	**Type**	**Source**	**Frequency**
questionnaire	Standardized questionnaire	The human operator's situation awareness can be measured using the Situation Awareness Global Assessment Technique (SAGAT) (Endsley, 1988) or similar.	In dependency of the particular experimental design

BASIC KPI INFORMATION			
ID	**Application Domain(s)**	**Name of KPI**	
KPI-HS-023	ATM, Power Grid, Railway	Human response time	
Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	31.10.2024	Duarte Dias (INESC TEC)	Creation of the document
1.0	03.03.2025	Duarte Dias (INESC TEC)	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	Human response time KPI evaluates time needed to react to AI advisory/information.
Objective(s)	This KPI assesses whether the inputs of the operators are according to their real psychophysiology. This can act as a verification methodology but also support the AI to adapt. This KPI contributes to evaluating Human-user experience of the AI-based assistant, as part of Task 4.3 evaluation objectives, and O3 main project objective.
Formula(s)	The time should be measured directly from user input and automatically by the system in background (dismiss a window when they feel satisfied after evaluating a scenario): • LOW less than 5 min, • MEDIUM 5-10 min, • HIGH more than 15 minutes. Then it is translated into % across the operator's multiple interactions with AI-generated solutions. This KPI is still under analysis on how it will be implemented. The objective is that it is transversal to all domains, but this means that an implementation will need to be done in each virtual environment. This implementation is still not defined and will need to be discussed with other Tasks/WPs
Unit of Measurement	LOW, MED, HIGH response time %

CALCULATION METHODOLOGY		
Baseline calculation methodology		
Step #	**Step description**	**Calculation**
1	Baseline data collection procedure – data collected outside the working environment, during a resting period .	See KPI calculation methodology
2	Standardized experimental stress procedure – Trier Social Stress Test - TSST, along with a 2-choice Reaction Time Task - that was successfully implemented by the research team in similar settings	See KPI calculation methodology
KPI calculation methodology		
KPI step #	**Step description**	**Calculation**
1	Operator monitoring in real-time during the different tests with collection of physiological data and computation of biomarkers in real-time to provide to the AI decision support system.	To be defined

Data collection			
Data ID	**Type**	**Source**	**Frequency**
n.a.	n.a.	n.a.	n.a.

BASIC KPI INFORMATION		
ID	**Application Domain(s)**	**Name of KPI**
KPI-HS-022	ATM, Power Grid, Railway	Human motivation

Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	06.02.2025	Patrick Zinsli (FHNW)	Creating document
1.0	03.03.2025	Patrick Zinsli (FHNW)	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	"Intrinsic motivation is defined as doing an activity for its inherent satisfaction rather than for some separable consequence. When intrinsically motivated, a person is moved to act for the fun or challenge entailed rather than because of external products, pressures, or rewards" (Ryan & Deci, 2000, p. 56).
Objective(s)	This KPI contributes to evaluating Human-user experience of the AI-based assistant, as part of Task 4.3 evaluation objectives, and O3 main project objective.
Formula(s)	As operationalized by the questionnaire (normally Likert-scales with several items that are rated on a scale of e.g. 1-5)
Unit of Measurement	Lickert-Scale or similar

CALCULATION METHODOLOGY		
Baseline calculation methodology		
Step #	**Step description**	**Module(s)**
1	Baseline is measured in an experimental setting with test persons in the control group (to see whether it differs from the experimental group). This requires a sufficient sample size.	Simulation and Testing Tools (Operations Testing Team)
KPI calculation methodology		
KPI step #	**Step description**	**Module(s)**
1	In dependency of the particular experimental design, the KPI is measured once or several times at defined points of time with test persons of the control group as well as of the experimental group by the means of a standardized questionnaire.	Simulation and Testing Tools (Operations Testing Team)

Data collection			
Data ID	**Type**	**Source**	**Frequency**
questionnaire	Adapted standardized questionnaire	The human operators' perceived internal work motivation can be measured by using the Job Diagnostic Survey (Hackman, et al., 1974) or something similar. The questionnaire must be adapted to the AI context (e.g., problem detection with AI assistance).	In dependency of the particular experimental design

BASIC KPI INFORMATION			
ID	**Application Domain(s)**	**Name of KPI**	
KPI-DS-015	ATM, Power Grid, Railway	Decision support satisfaction	
Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	07.03.2025	Kostiantyn Kucher (LiU)	Initial version
1.0	08.03.2025	Kostiantyn Kucher (LiU)	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	This KPI represents human operators' self-reported satisfaction with the system's support for their decision-making process when working with the AI assistant measured with a questionnaire.
Objective(s)	This KPI contributes to evaluating Human user experience of the AI-based assistant, as part of Task 4.3 evaluation objectives, and O3 main project objective. It is also relevant to protocols and concepts defined in D1.1 such as "Decision support for the human operator", "Decision support satisfaction".
Formula(s)	As operationalized by the questionnaire (normally Likert-scales with several items that are rated on a scale of e.g. 1–5 or 1–7).
Unit of Measurement	Ordinal data response on a Likert scale (or potentially a similar response on an interval scale)

CALCULATION METHODOLOGY		
Baseline calculation methodology		
Step #	**Step description**	**Calculation**
1	An existing questionnaire such as ATAI (Sindermann, et al., 2021) can be used to capture the self-reported pre-study attitude towards the AI assistant as an indirect baseline.	Simulation and Testing Tools (Operations Testing Team)
KPI calculation methodology		
KPI step #	**Step description**	**Calculation**
1	The KPI is measured by the means of a questionnaire comprising one or several questions on a Likert scale. The basic version of the questionnaire could consist of one or several questions / statements to be rated such as "I am satisfied with the support for my decision-making process provided by the AI assistant". Further questions could be adapted from the items relevant to decision-making support from the Human-Computer Trust questionnaire (Madsen, et al., 2000) as well as the Likert-scale questions on AI and explanation usefulness (Bansal, et al., 2021). The contents of the questionnaire should be adjusted and initially tested according to the particular experimental design.	Simulation and Testing Tools (Operations Testing Team)

Data collection			
Data ID	**Type**	**Source**	**Frequency**
questionnaire	Questionnaire with one or several questions on a Likert scale	A novel questionnaire designed specifically for this study or adapted from the existing work, such as the questions relevant to decision- making from the Human-Computer Trust questionnaire (Madsen, et al., 2000) as well as the Likert-scale questions on AI and explanation usefulness (Bansal, et al., 2021).	In dependency of the particular experimental design
questionnaire2	Questionnaire with one or several questions on a Likert scale	An auxiliary questionnaire can be considered to gather accompanying information, e.g., one of the standard perceived usability/satisfaction questionnaires such as SUS (Brooke, 1996) or UMUX-LITE (Lewis, et al., 2013).	In dependency of the particular experimental design

AI-HUMAN LEARNING CURVE

BASIC KPI INFORMATION		
ID	**Application Domain(s)**	**Name of KPI**
KPI-AS-006	ATM, Power Grid, Railway	AI co-learning capability

Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	07.03.2025	Kostiantyn Kucher (LiU)	Initial version
1.0	08.03.2025	Kostiantyn Kucher (LiU)	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	This KPI represents human operators' self-reported assessment of the AI ability to adapt to the operators' preferences measured with a questionnaire.
Objective(s)	This KPI contributes to evaluating AI-human learning curves of the AI-based assistant, as part of Task 4.3 evaluation objectives, and O3 main project objective. It is also relevant to protocols and concepts defined in D1.1 such as "AI co-learning capability".
Formula(s)	As operationalized by the questionnaire (normally Likert-scales with several items that are rated on a scale of e.g. 1–5 or 1–7).
Unit of Measurement	Ordinal data response on a Likert scale (or potentially a similar response on an interval scale)

CALCULATION METHODOLOGY		
Baseline calculation methodology		
Step #	**Step description**	**Calculation**
1	An existing questionnaire such as ATAI (Sindermann, et al., 2021) can be used to capture the self-reported pre-study attitude towards the AI assistant as an indirect baseline.	Simulation and Testing Tools (Operations Testing Team)
KPI calculation methodology		
KPI step #	**Step description**	**Calculation**
1	The KPI is measured by the means of a questionnaire comprising one or several questions on a Likert scale. The basic version of the questionnaire could consist of one or several questions / statements to be rated such as "The AI assistant has been able to adapt to the most important of my preferences" and "The AI assistant has been able to adapt to all of my preferences". Further questions could be adapted from the human-robot collaboration questionnaire (Nikolaidis, et al., 2017), such as "The AI assistant perceived accurately what my goals are"; the subjective fluency metric scales for human-machine systems (Hoffman, et al., 2018); and the questionnaire on perception of role and effects of adaptive AI teammates (Hauptman, et al., 2023). The contents of the questionnaire should be adjusted and initially tested according to the particular experimental design.	Simulation and Testing Tools (Operations Testing Team)

Data collection			
Data ID	**Type**	**Source**	**Frequency**
$questionnaire$	Questionnaire with one or several questions on a Likert scale	A novel questionnaire designed specifically for this study or adapted from the existing work, including the human-robot collaboration questionnaire (Nikolaidis, et al., 2017), the subjective fluency metric scales for human-machine systems (Hoffman, et al., 2018) and the role and effects of adaptive AI teammates (Hauptman, et al., 2023).	In dependency of the particular experimental design
$questionnaire2$	Questionnaire with one or several questions on a Likert scale	An auxiliary questionnaire can be considered to gather accompanying information, such as human-AI collaboration questionnaire (Du, et al., 2025) or the ASA Questionnaire (Fitrianie, et al., 2022).	In dependency of the particular experimental design

BASIC KPI INFORMATION		
ID	**Application Domain(s)**	**Name of KPI**
KPI-HS-021	ATM, Power Grid, Railway	Human learning

Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	06.02.2025	Patrick Zinsli (FHNW)	Creating document
1.0	03.03.2025	Patrick Zinsli (FHNW)	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	Human learning is a complex process that leads to lasting changes in humans, influencing their perceptions of the world and their interactions with it across physical, psychological, and social dimensions. It is fundamentally shaped by the ongoing, interactive relationship between the learner's characteristics and the learning content, all situated within the specific environmental context of time and place and the continuity over time.
Objective(s)	This KPI contributes to evaluating AI-human learning curves of the AI-based assistant, as part of Task 4.3 evaluation objectives, and O3 main project objective.
Formula(s)	As operationalized by the questionnaire (normally Likert-scales with several items that are rated on a scale of e.g. 1-5)
Unit of Measurement	Lickert-Scale or similar

CALCULATION METHODOLOGY

Baseline calculation methodology

Step #	Step description	Calculation
1	Baseline is measured in an experimental setting with test persons in the control group (to see whether it differs from the experimental group). This requires a sufficient sample size.	Simulation and Testing Tools (Operations Testing Team)

KPI calculation methodology

KPI step #	Step description	Calculation
1	In dependency of the particular experimental design, the KPI is measured once or several times at defined points of time with test persons of the control group as well as of the experimental group by the means of a standardized questionnaire.	Simulation and Testing Tools (Operations Testing Team)

Data collection

Data ID	Type	Source	Frequency
Questionnaire	Questionnaire	The human operators' perceived learning opportunities working with the AI-based system can be measured using the task-based workplace learning scale (Nikolova, et al., 2014) or something similar. The questionnaire needs to be adapted to the AI context.	In dependency of the particular experimental design

AI-HUMAN TASK ALLOCATION BALANCE

BASIC KPI INFORMATION		
ID	**Application Domain(s)**	**Name of KPI**
KPI-HS-018	ATM, Power Grid, Railway	Human control/autonomy over the process

Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	07.03.2025	Kostiantyn Kucher (LiU)	Initial version
1.0	08.03.2025	Kostiantyn Kucher (LiU)	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	This KPI represents human operators' perceived autonomy over the process when working with the AI assistant measured with a questionnaire.
Objective(s)	This KPI contributes to evaluating AI-human task allocation balance of the AI-based assistant, as part of Task 4.3 evaluation objectives, and O3 main project objective. It is also relevant to protocols and concepts defined in D1.1 such as "Decision support for the human operator", "Human Agency and Oversight", "Human control/autonomy over the process".
Formula(s)	As operationalized by the questionnaire (normally Likert-scales with several items that are rated on a scale of e.g. 1–5 or 1–7).
Unit of Measurement	Ordinal data response on a Likert scale (or potentially a similar response on an interval scale)

CALCULATION METHODOLOGY		
Baseline calculation methodology		
Step #	**Step description**	**Calculation**
1	An existing questionnaire such as SoAS (Tapal, et al., 2017) or HILDA (Fraser, 2010) can be used to capture the self-reported perception of several dimensions of agency as an indirect baseline.	Simulation and Testing Tools (Operations Testing Team)
KPI calculation methodology		
KPI step #	**Step description**	**Calculation**
1	The KPI is measured by the means of the Work Design Questionnaire (Morgeson, et al., 2006), the human-AI collaboration questionnaire (Du, et al., 2025), or similar comprising several questions on a Likert scale. The contents of the questionnaire should be adjusted to the AI context (e.g., problem detection with AI assistance) and initially tested according to the particular experimental design.	Simulation and Testing Tools (Operations Testing Team)

Data collection			
Data ID	**Type**	**Source**	**Frequency**
questionnaire	Questionnaire with one or several questions on a Likert scale	A novel questionnaire designed specifically for this study or adapted from the existing work, such as WDQ (Morgeson, et al., 2006) or human-AI collaboration questionnaire (Du, et al., 2025).	In dependency of the particular experimental design

$questionnaire2$	Questionnaire with one or several questions on a Likert scale	An auxiliary questionnaire can be considered to gather accompanying information, such as the IE-4 questionnaire for Locus of Control (Nießen, et al., 2022) or the ASA Questionnaire (Fitrianie, et al., 2022).	In dependency of the particular experimental design

BASIC KPI INFORMATION			
ID	**Application Domain(s)**	**Name of KPI**	
KPI-IS-041	ATM, Power Grid, Railway	Impact on the workload	
Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	31.10.2024	Duarte Dias (INESC TEC)	Creation of the document
1.0	03.03.2025	Duarte Dias (INESC TEC)	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	Impact on the workload KPI assesses operators' perception of the system impact on their workload (either positive or negative)
Objective(s)	This KPI compares if the inputs of the operators are according to their real psychophysiology. This can act as a verification methodology but also support the AI to adapt. This KPI will be analyzed together with the "Workload" KPI-WS-040. This KPI contributes to evaluating AI-human task allocation balance of the AI-based assistant, as part of Task 4.3 evaluation objectives, and O3 main project objective.
Formula(s)	It is measured directly from user input using a 7-point Likert scale: • From 1 (Huge Increase in workload) • To 7 (Huge decrease of workload). This KPI is still under analysis on how it will be implemented. If with a single manual questionnaire or with a pop-up in the dashboard.
Unit of Measurement	Value between 1 and 7

CALCULATION METHODOLOGY		
Baseline calculation methodology		
Step #	**Step description**	**Calculation**
1	Baseline data collection procedure – data collected outside the working environment, during a resting period .	See KPI calculation methodology
2	Standardized experimental stress procedure – Trier Social Stress Test - TSST, along with a 2-choice Reaction Time Task - that was successfully implemented by the research team in similar settings	See KPI calculation methodology
KPI calculation methodology		
KPI step #	**Step description**	**Calculation**
1	Operator monitoring in real-time during the different tests with collection of physiological data and computation of biomarkers in real-time to provide to the AI decision support system.	To be defined

Data collection			
Data ID	**Type**	**Source**	**Frequency**
n.a.	n.a.	n.a.	n.a.

LONG-TERM CONSEQUENCES OF AI ASSISTANTS

BASIC KPI INFORMATION		
ID	**Application Domain(s)**	**Name of KPI**
KPI-RS-091	ATM, Power Grid, Railway	Reflection on operator trust

Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	05.03.2025	Kostiantyn Kucher (LiU)	Initial version
1.0	08.03.2025	Kostiantyn Kucher (LiU)	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	This KPI represents self-reported human operators' perception of the changes in their trust for the AI assistant over time (increased/decreased) on a Likert scale.
Objective(s)	This KPI contributes to evaluating Long-term consequences of AI assistants of the AI-based assistant, as part of Task 4.3 evaluation objectives, and O3 main project objective. It is also relevant to protocols and concepts defined in D1.1 such as "Transparency", "Human Agency and Oversight", "Credibility and Intimacy". Furthermore, it is also relevant to the overall project KPI-ET-7 "% of acceptance of human operators regarding AI4REALNET solutions".
Formula(s)	As operationalized by the questionnaire (normally Likert-scales with several items that are rated on a scale of e.g. 1–5 or 1–7).
Unit of Measurement	Ordinal data response on a Likert scale (or potentially a similar response on an interval scale)

CALCULATION METHODOLOGY		
Baseline calculation methodology		
Step #	**Step description**	**Calculation**
1	The values of KPI-TS-039 "Trust towards the AI tool" can be used as an indirect baseline for analyzing the changes over time for a target audience sample (including ideally the same, but potentially also different subjects/participants). Alternatively, a variation of the questionnaire used for this KPI (with questions adapted from reflection over changes to the attitude at current moment in time) could also be employed. Additionally, an existing questionnaire such as ATAI (Sindermann, et al., 2021) can be used to capture the self-reported pre-study attitude towards the AI assistant as an indirect baseline.	Simulation and Testing Tools (Operations Testing Team)
KPI calculation methodology		
KPI step #	**Step description**	**Calculation**
1	The KPI is measured by the means of a questionnaire comprising one or several questions on a Likert scale. The basic version of the questionnaire could consist of two questions / statements to be rated such as "My trust for the AI assistant has increased over time" and "My trust for the AI assistant has decreased over time" to capture the direction and amplitude of the change perceived by the respondents. The contents of the questionnaire should be adjusted and initially tested according to the particular experimental design.	Simulation and Testing Tools (Operations Testing Team)

Data collection			
Data ID	**Type**	**Source**	**Frequency**
questionnaire	Questionnaire with one or several questions on a Likert scale	A novel questionnaire designed specifically for this study or adapted from the existing work such as the Scale for XAI (Hoffman, et al., 2018), ATAI (Sindermann, et al., 2021), or similar. The dimension of decreasing trust can be informed by the prior work on distrust in AI (Peters, et al., 2023).	In dependency of the particular experimental design

BASIC KPI INFORMATION			
ID	**Application Domain(s)**	**Name of KPI**	
KPI-RS-092	ATM, Power Grid, Railway	Reflection on operator agency	
Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	05.03.2025	Kostiantyn Kucher (LiU)	Initial version
1.0	08.03.2025	Kostiantyn Kucher (LiU)	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	This KPI represents self-reported human operators' perception of the changes in their agency working with the AI assistant over time (increased/decreased) on a Likert scale.
Objective(s)	This KPI contributes to evaluating Long-term consequences of AI assistants of the AI-based assistant, as part of Task 4.3 evaluation objectives, and O3 main project objective. It is also relevant to protocols and concepts defined in D1.1 such as "Transparency", "Decision support for the human operator", "Human Agency and Oversight".
Formula(s)	As operationalized by the questionnaire (normally Likert-scales with several items that are rated on a scale of e.g. 1–5 or 1–7).
Unit of Measurement	Ordinal data response on a Likert scale (or potentially a similar response on an interval scale)

CALCULATION METHODOLOGY		
Baseline calculation methodology		
Step #	**Step description**	**Calculation**
1	An existing questionnaire such as SoAS (Tapal, et al., 2017) or HILDA (Fraser, 2010) can be used to capture the self-reported perception of several dimensions of agency as an indirect baseline for analyzing the changes over time for a target audience sample (including ideally the same, but potentially also different subjects/participants). A variation of the questionnaire used for this KPI (with questions adapted from reflection over changes to the attitude at current moment in time) could also be employed.	Simulation and Testing Tools (Operations Testing Team)
KPI calculation methodology		
KPI step #	**Step description**	**Calculation**
1	The KPI is measured by the means of a questionnaire comprising one or several questions on a Likert scale. The basic version of the questionnaire could consist of two questions / statements to be rated such as "My sense of agency during interactions with the AI assistant has increased over time" and "My sense of agency during interactions with the AI assistant has decreased over time" to capture the direction and amplitude of the change perceived by the respondents. Further questions could be adapted from the Work Design Questionnaire (Morgeson, et al., 2006) or the human-AI collaboration questionnaire (Du, et al., 2025). The contents of the questionnaire should be adjusted and initially tested according to the particular experimental design.	Simulation and Testing Tools (Operations Testing Team)

Data collection			
Data ID	**Type**	**Source**	**Frequency**
questionnaire	Questionnaire with one or several questions on a Likert scale	A novel questionnaire designed specifically for this study or adapted from the existing work, such as the SoAS questionnaire (Tapal, et al., 2017), WDQ (Morgeson, et al., 2006) or human-AI collaboration questionnaire (Du, et al., 2025).	In dependency of the particular experimental design
questionnaire2	Questionnaire with one or several questions on a Likert scale	An auxiliary questionnaire can be considered to gather accompanying information, such as the IE-4 questionnaire for Locus of Control (Nießen, et al., 2022) or the ASA Questionnaire (Fitrianie, et al., 2022).	In dependency of the particular experimental design

BASIC KPI INFORMATION			
ID	**Application Domain(s)**	**Name of KPI**	
KPI-RS-093	ATM, Power Grid, Railway	Reflection on operator deskilling	
Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	05.03.2025	Kostiantyn Kucher (LiU)	Initial version
1.0	08.03.2025	Kostiantyn Kucher (LiU)	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	This KPI represents self-reported human operators' perception of the changes in their own skills working with the AI assistant over time (increased/decreased) on a Likert scale.
Objective(s)	This KPI contributes to evaluating Long-term consequences of AI assistants of the AI-based assistant, as part of Task 4.3 evaluation objectives, and O3 main project objective. It is also relevant to protocols and concepts defined in D1.1 such as "Mitigate de-skilling in the human operators".
Formula(s)	As operationalized by the questionnaire (normally Likert-scales with several items that are rated on a scale of e.g. 1–5 or 1–7).
Unit of Measurement	Ordinal data response on a Likert scale (or potentially a similar response on an interval scale)

CALCULATION METHODOLOGY		
Baseline calculation methodology		
Step #	**Step description**	**Calculation**
1	An existing questionnaire such as HILDA (Fraser, 2010) can be used to capture the self-reported perception of several dimensions of skill at a given point (including skill-intensity, autonomy/control, and substantive complexity) as an indirect baseline for analyzing the changes over time for a target audience sample (including ideally the same, but potentially also different subjects/participants). A variation of the questionnaire used for this KPI (with questions adapted from reflection over changes to the attitude at current moment in time) could also be employed.	Simulation and Testing Tools (Operations Testing Team)
KPI calculation methodology		
KPI step #	**Step description**	**Calculation**
1	The KPI is measured by the means of a questionnaire comprising one or several questions on a Likert scale. The basic version of the questionnaire could consist of three questions / statements to be rated such as "My skills have increased over time", "My skills have decreased over time", and "Working with the AI assistant was the reason for the change of my skills over time" to capture the direction and amplitude of the change perceived by the respondents. The contents of the questionnaire should be adjusted and initially tested according to the particular experimental design.	Simulation and Testing Tools (Operations Testing Team)

Data collection			
Data ID	**Type**	**Source**	**Frequency**
questionnaire	Questionnaire with one or several questions on a Likert scale	A novel questionnaire designed specifically for this study or adapted from the existing work such as HILDA (Fraser, 2010).	In dependency of the particular experimental design
questionnaire2	Questionnaire with one or several questions on a Likert scale	An auxiliary questionnaire can be considered to gather accompanying information, such as the AI-Acceptance model (KIAM) (Scheuer, 2020).	In dependency of the particular experimental design

BASIC KPI INFORMATION			
ID	**Application Domain(s)**	**Name of KPI**	
KPI-RS-094	ATM, Power Grid, Railway	Reflection on over-reliance	
Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	05.03.2025	Kostiantyn Kucher (LiU)	Initial version
1.0	08.03.2025	Kostiantyn Kucher (LiU)	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	This KPI represents self-reported human operators' perception of their potential over-reliance on the AI assistant on a Likert scale.
Objective(s)	This KPI contributes to evaluating Long-term consequences of AI assistants of the AI-based assistant, as part of Task 4.3 evaluation objectives, and O3 main project objective. It is also relevant to protocols and concepts defined in D1.1 such as "Mitigate addictive behavior from humans".
Formula(s)	As operationalized by the questionnaire (normally Likert-scales with several items that are rated on a scale of e.g. 1–5 or 1–7).
Unit of Measurement	Ordinal data response on a Likert scale (or potentially a similar response on an interval scale)

CALCULATION METHODOLOGY		
Baseline calculation methodology		
Step #	**Step description**	**Calculation**
1	The values of KPI-TS-039 "Trust towards the AI tool" can be used as an indirect baseline for assessing the underlying trust for a target audience sample (including ideally the same, but potentially also different subjects/participants). Alternatively, a variation of the questionnaire used for this KPI (with questions adapted from reflection over changes to the attitude at current moment in time) could also be employed. Additionally, an existing questionnaire such as ATAI (Sindermann, et al., 2021) can be used to capture the self-reported pre-study attitude towards the AI assistant as an indirect baseline.	Simulation and Testing Tools (Operations Testing Team)
KPI calculation methodology		
KPI step #	**Step description**	**Calculation**
1	The KPI is measured by the means of a questionnaire comprising one or several questions on a Likert scale. The basic version of the questionnaire could consist of two questions / statements to be rated such as "I started relying on the AI assistant and following all of its suggestions without reservations over time" and "The design of the AI assistant discourages audit and interventions from the operator" to capture the direction and amplitude of the change perceived by the respondents. Further questions could be adapted from the items relevant to decision-making support from the Human-Computer Trust questionnaire (Madsen, et al., 2000). The contents of the questionnaire should be adjusted and initially tested according to the particular experimental design.	Simulation and Testing Tools (Operations Testing Team)

Data collection			
Data ID	**Type**	**Source**	**Frequency**
questionnaire	Questionnaire with one or several questions on a Likert scale	A novel questionnaire designed specifically for this study or adapted from the existing work such as the Human-Computer Trust questionnaire (Madsen, et al., 2000), ATAI (Sindermann, et al., 2021), or similar. The dimension of overreliance can be informed by the prior work on distrust in AI (Peters, et al., 2023).	In dependency of the particular experimental design

BASIC KPI INFORMATION			
ID	**Application Domain(s)**	**Name of KPI**	
KPI-RS-094	ATM, Power Grid, Railway	Reflection on additional training	
Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	05.03.2025	Kostiantyn Kucher (LiU)	Initial version
1.0	08.03.2025	Kostiantyn Kucher (LiU)	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	This KPI represents self-reported human operators' perception of the additional training necessary to adopt the AI assistant on a Likert scale.
Objective(s)	This KPI contributes to evaluating Long-term consequences of AI assistants of the AI-based assistant, as part of Task 4.3 evaluation objectives, and O3 main project objective. It is also relevant to protocols and concepts defined in D1.1 such as "Additional training about AI for human operators" and "Societal and Environmental Well-being".
Formula(s)	As operationalized by the questionnaire (normally Likert-scales with several items that are rated on a scale of e.g. 1–5 or 1–7).
Unit of Measurement	Ordinal data response on a Likert scale (or potentially a similar response on an interval scale)

CALCULATION METHODOLOGY		
Baseline calculation methodology		
Step #	**Step description**	**Calculation**
1	The values of KPI-HS-021 "Human learning" can be used as an indirect baseline for analyzing the changes over time for a target audience sample (including ideally the same, but potentially also different subjects/participants). Alternatively, a variation of the questionnaire used for this KPI (with questions adapted from reflection over changes to the attitude at current moment in time) could also be employed.	Simulation and Testing Tools (Operations Testing Team)
KPI calculation methodology		
KPI step #	**Step description**	**Calculation**
1	The KPI is measured by the means of a questionnaire comprising one or several questions on a Likert scale. The basic version of the questionnaire could consist of several questions / statements to be rated such as "I feel that I need to undergo additional training in order to work with the AI assistant" and "I feel that the additional training in order to work with the AI assistant will demand considerable time and effort". The contents of the questionnaire should be adjusted and initially tested according to the particular experimental design.	Simulation and Testing Tools (Operations Testing Team)

Data collection			
Data ID	**Type**	**Source**	**Frequency**
questionnaire	Questionnaire with one or several questions on a Likert scale	A novel questionnaire designed specifically for this study or adapted from the existing work, including the dimensions of access to knowledge (Bedué, et al., 2022).	In dependency of the particular experimental design

$questionnaire2$	Questionnaire with one or several questions on a Likert scale	An auxiliary questionnaire can be considered to gather accompanying information, such as the IDS scale for the intensification of job demands (Kubicek, et al., 2015) or the AI anxiety scale (Wan, et al., 2022).	In dependency of the particular experimental design

BASIC KPI INFORMATION			
ID	**Application Domain(s)**	**Name of KPI**	
KPI-RS-096	ATM, Power Grid, Railway	Reflection on biases	
Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	05.03.2025	Kostiantyn Kucher (LiU)	Initial version
1.0	08.03.2025	Kostiantyn Kucher (LiU)	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	This KPI represents self-reported human operators' perception of biased decisions potentially produced by the AI assistant with respect to gender/ethnicity/age or commercial interests on a Likert scale.
Objective(s)	This KPI contributes to evaluating Long-term consequences of AI assistants of the AI-based assistant, as part of Task 4.3 evaluation objectives, and O3 main project objective. It is also relevant to protocols and concepts defined in D1.1 such as "Diversity, Non-discrimination, and Fairness".
Formula(s)	As operationalized by the questionnaire (normally Likert-scales with several items that are rated on a scale of e.g. 1–5 or 1–7).
Unit of Measurement	Ordinal data response on a Likert scale (or potentially a similar response on an interval scale)

CALCULATION METHODOLOGY		
Baseline calculation methodology		
Step #	**Step description**	**Calculation**
1	The values of KPI-TS-038 "Trust in AI solutions score" and KPI-TS-039 "Trust towards the AI tool" can be used as an indirect baseline for analyzing the changes over time for a target audience sample (including ideally the same, but potentially also different subjects/participants). Alternatively, a variation of the questionnaire used for this KPI (with questions adapted from reflection over changes to the attitude at current moment in time) could also be employed.	Simulation and Testing Tools (Operations Testing Team)
KPI calculation methodology		
KPI step #	**Step description**	**Calculation**
1	The KPI is measured by the means of a questionnaire comprising one or several questions on a Likert scale. The basic version of the questionnaire could consist of several questions / statements to be rated such as "The solutions proposed by the AI assistant are affected by gender/ethnicity/age of person(s) mentioned in the data", "The solutions proposed by the AI assistant are affected by the organizations/affiliations mentioned in the data", and "The solutions proposed by the AI assistant are affected by gender/ethnicity/age of the operator" to capture the opinion of the respondents. The contents of the questionnaire should be adjusted and initially tested according to the particular experimental design.	Simulation and Testing Tools (Operations Testing Team)

Data collection			
Data ID	**Type**	**Source**	**Frequency**
questionnaire	Questionnaire with one or several questions on a Likert scale	A novel questionnaire designed specifically for this study or adapted from the existing work, including the trust dimensions relevant to AI adoption (Bedué, et al., 2022) as well as potential sources of bias in AI (Lee, et al., 2021); (Schwartz, et al., 2022) and perception of such biases (Kim, 2025).	In dependency of the particular experimental design

BASIC KPI INFORMATION			
ID	**Application Domain(s)**	**Name of KPI**	
KPI-RS-097	ATM, Power Grid, Railway	Predicted long-term adoption	
Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	05.03.2025	Kostiantyn Kucher	Initial version
1.0	08.03.2025	Kostiantyn Kucher	Finalization for D4.1

SCOPE AND OBJECTIVES OF KPI	
Description	This KPI represents predicted adoption of the AI assistant by users, stakeholders, or experts on a Likert scale.
Objective(s)	This KPI contributes to evaluating Long-term consequences of AI assistants of the AI-based assistant, as part of Task 4.3 evaluation objectives, and O3 main project objective. It is also relevant to protocols and concepts defined in D1.1 such as “Human Agency and Oversight”, “Societal and Environmental Well-being”.
Formula(s)	As operationalized by the questionnaire (normally Likert-scales with several items that are rated on a scale of e.g. 1–5 or 1–7).
Unit of Measurement	Ordinal data response on a Likert scale (or potentially a similar response on an interval scale)

CALCULATION METHODOLOGY		
Baseline calculation methodology		
Step #	**Step description**	**Calculation**
1	The values of KPI-AS-002 “Acceptance” can be used as an indirect baseline for analyzing the changes over time for a target audience sample (including ideally the same, but potentially also different subjects/participants). Alternatively, a variation of the questionnaire used for this KPI (with questions adapted from reflection over changes to the attitude at current moment in time) could be employed. Additionally, an existing questionnaire such as ATAI (Sindermann, et al., 2021) can be used to capture the self-reported pre-study attitude towards the AI assistant as an indirect baseline.	Simulation and Testing Tools (Operations Testing Team)
KPI calculation methodology		
KPI step #	**Step description**	**Calculation**
1	The KPI is measured by the means of a questionnaire comprising one or several questions on a Likert scale. The basic version of the questionnaire could consist of several questions / statements to be rated such as “The proposed AI assistant is likely to be adopted by end-users in the long term” and “The overall methodology and technology (not necessarily the proposed AI assistant) is likely to be adopted by end-users in the long term”. Further questions could be adapted from the questionnaires on perception of role and effects of adaptive AI teammates (Hauptman, et al., 2023) and AI adoption (Bedué, et al., 2022). The contents of the questionnaire should be adjusted and initially tested according to the particular experimental design.	Simulation and Testing Tools (Operations Testing Team)

Data collection			
Data ID	**Type**	**Source**	**Frequency**
$questionnaire$	Questionnaire with one or several questions on a Likert scale	A novel questionnaire designed specifically for this study or adapted from the existing work, including the role and effects of adaptive AI teammates (Hauptman, et al., 2023) and trust dimensions relevant to AI adoption (Bedué, et al., 2022).	In dependency of the particular experimental design
$questionnaire2$	Questionnaire with one or several questions on a Likert scale	An auxiliary questionnaire can be considered to gather accompanying information, such as the IDS scale for the intensification of job demands (Kubicek, et al., 2015) or the AI anxiety scale (Wan, et al., 2022).	In dependency of the particular experimental design

ANNEX 2 – OPERATIONAL SCENARIOS

Page intentionally left blank.

SCENARIO TEMPLATE

This template has been used to describe all domain specific operational scenarios.

1 Name of the Operational Scenario

ID and name shall be consistent with the D4.1_Evaluation and test protocol document.

ID	*Application Domain(s)*	*Name of Operational Scenario*
Text	Text	Text

2 Version management

Version Management			
Version No.	*Date*	*Name of Author(s)*	*Changes*
0.1	DD.MM.YYYY	Text	Text

3 Narrative of the Operational Scenario

References shall be consistent with the use case documents.

Description of Operational Scenario
Use case
Name of the use case corresponding to the scenario (linked back to use case document) Text
Complete description
Operational scenarios describe examples of how users/operators/maintainers will interact with the system in important contexts of use. The scenarios are described for an activity or a series of activities of business processes supported by the system. Description should be taken from the use case document: the description should be concise and stick to the description of the testing scenario to avoid duplication of information. If needed, put reference to longer use case description. Text

4 Detailed steps

Describe in detail the operational scenario, starting from what has been described in the corresponding use case.
A focus shall be made on the post-conditions, which are more important here: i.e., which is the final state of the system (or result) in each scenario.

Operational Scenario		
Type	*No.*	*Description*
Pre-condition, step, post-condition	1, 2, 3, etc.	Detailed description
Text	Text	Text

5 Instances of the Operational Scenario

Describe concrete instances/distribution of instances for this operational scenario.
For each instance specify which dataset is used

Instances	
Name	*Description*
Short name. Text	Detailed description. Text

ATM SCENARIOS

Page intentionally left blank.

1 Name of the Operational Scenario

ID	*Application Domain(s)*	*Name of Operational Scenario*
OPSCE-UC1.A-1-012	ATM	Airspace sectorization assistant – Adverse weather conditions

2 Version management

Version Management			
Version No.	*Date*	*Name of Author(s)*	*Changes*
0.1	13.01.2025	Cristina Félix (NAV), João Soares (NAV)	Initial Version
0.2	23.01.2025	Cristina Félix (NAV), João Soares (NAV), Joaquim Geraldes (NAV)	First revision
1.0	30.01.2025	Cristina Félix (NAV), João Soares (NAV), Joaquim Geraldes (NAV)	Final version

3 Narrative of the Operational Scenario

Description of Operational Scenario
Use case
UC1.ATM - Airspace sectorisation
Complete description
This operational scenario is based on the airspace sectorization which involves retrieving and integrating several data information sources that are often gathered from different (digital) platforms, such as: 1. Expected traffic counts (available from EUROCONTROL NM) 2. Air-ground and coordination message count 3. Weather Information (METEO fore- and now casts) 4. Airspace Reservations (e.g., military airspace, temporary 'no-fly' zones) 5. Coordination Complexity (e.g., between area and arrival controllers) 6. Terminal Area Complexity (e.g., weather-related airport capacity limitations) 7. Equipment issues (e.g., communication issues between pilots and air traffic controllers) 8. ATCO staff schedules (depending on traffic forecasts demands) This scenario focuses on a new sectorization plan in response to adverse weather conditions (e.g. Volcanic ashes). The AI-based system detects the perturbation, recalculates new sectorization plans and applies an optimized solution to avoid the affected airspace. **Complete Description**: 1. **Trigger**: A portion of airspace becomes unavailable due to volcanic ashes from a volcanic eruption. 2. **System Reaction**: The AI system detects the perturbation, assesses its impact, and proposes new sectorization plans. 3. **Operator Role**: The supervisor monitors the AI decisions, validates, or adapts the solution, and decides. 4. **Goal**: Optimal sectorization readjustments to minimize delays.

4 Detailed steps

Operational Scenario		
Type	***No.***	***Description***
Pre-condition	1	Initial state of the system
Step	2	Staff manager looks at estimated traffic counts, and operational conditions, and using his experience decides the sectorization plan. He/She looks at available ATCO staff during a shift, selects a maximum time horizon for a sector plan and enters that information into the system.
Step	3	The staff manager requests an initial sectorisation plan from the AI assistant. This plan includes portraying a horizontal and vertical sector layout on a map and/or secondary interface, timeline showing ATCO staff occupancy per sector, time slider enabling the staff manager to preview changes in sectorisation plans on a map. The predicted state of the system in terms of traffic movements and weather condition (e.g., wind) is also displayed and responsive to the time slider.
Step	4	The AI assistant may propose several alternative sector plans, each with a different probability value (based on historical data) and robustness score depending on available ATCO staff, fluctuations in predicted traffic load, and uncertainty in weather forecasts. Using the time slider, the staff manager can evaluate the probability and robustness scores for different times within the maximum look-ahead time horizon.
Step	5	The staff manager interacts with the suggested sector plan in one of the following ways: 1) accept the top-rated AI suggestion and implement it; 2) nudge the AI suggestions by making small changes (e.g., one merge or split); 3) revise large sections of the plan (e.g., revise multiple sectorisation events across various time horizons).
Post-condition	6	The AI assistant monitors changes in predicted system and environmental states. When updated information deviates from the information and data that was used for the implemented sector plan, the AI assistant issues an alert, triggering the staff manager to go back to Step 2.

Operational Scenario		
Type	***No.***	***Description***
Pre-condition	1	The system operates with normal conditions.
Step	2	An unexpected weather event (volcanic ashes).
Step	3	The AI system determines the need for new sectorization plan through real-time monitoring of weather information and air traffic flight execution.
Step	4	The AI system calculates an alternative sectorization plan considering sector load balance and affected area.
Step	5	The proposed solution is displayed to the ATC supervisor for review.
Step	6	The ATC supervisor either approves the new plan or intervenes to improve upon the AI system advice.
Post-condition	7	The adjusted plan is executed, and delays are minimized to the greatest extent possible.

5 Instances of the Operational Scenario

Instances	
Name	***Description***
Weather perturbations	Some weather conditions can impact the normal functioning of airspaces such as volcanic ashes, severe winter weather, thunderstorms.

1 Name of the Operational Scenario

ID	Application Domain(s)	Name of Operational Scenario
OPSCE-UC2.A-1-011	ATM	Flow & Airspace management assistant – Military Reservation

2 Version management

Version Management			
Version No.	Date	Name of Author(s)	Changes
0.1	13.01.2025	Cristina Félix (NAV), João Soares (NAV)	Initial Version
0.2	23.01.2025	Cristina Félix (NAV), João Soares (NAV), Joaquim Geraldes (NAV)	First revision
1.0	30.01.2025	Cristina Félix (NAV), João Soares (NAV), Joaquim Geraldes (NAV)	Final version

3 Narrative of the Operational Scenario

Description of Operational Scenario
Use case
UC2.ATM - Flow & Airspace management
Complete description
This operational scenario is based on the airspace management which involves retrieving and integrating several data information sources that are often gathered from different (digital) platforms, such as: 9. Expected traffic counts (available from EUROCONTROL NM) 10. Air-ground and coordination message count 11. Weather Information (METEO fore- and now casts) 12. Airspace Reservations (e.g., military airspace, temporary 'no-fly' zones) 13. Coordination Complexity (e.g., between area and arrival controllers) 14. Terminal Area Complexity (e.g., weather-related airport capacity limitations) 15. Equipment issues (e.g., communication issues between pilots and air traffic controllers) 16. ATCO staff schedules (depending on traffic forecasts demands) This scenario focuses on a new sectorization plan and/or routing in response to a sudden airspace reservation. The AI-based system detects the disruption, recalculates flight routes and the need for a new sectorization plan (if possible) and new traffic load forecasts and applies an optimized solution to minimize delays while maintaining operational constraints. **Complete Description**: 5. **Trigger**: A portion of airspace becomes unavailable due to military reservations. 6. **System Reaction**: The AI system detects the disruption, assesses its impact, and proposes alternative a/c course deviations and/or different sectors. 7. **Operator Role**: The supervisor monitors the AI decisions, validates, or adapts the solution, and decides. 8. **Goal**: Optimal sector and a/c course deviations readjustments to minimize delays.

4 Detailed steps

Operational Scenario		
Type	***No.***	***Description***
Pre-condition	1	Initial state of the system
Step	2	Staff manager looks at estimated traffic counts, and operational conditions, and using his experience decides the sectorization plan. He/She looks at available ATCO staff during a shift, selects a maximum time horizon for a sector plan and enters that information into the system.
Step	3	The staff manager requests an initial sectorisation plan from the AI assistant. This plan includes portraying a horizontal and vertical sector layout on a map and/or secondary interface, timeline showing ATCO staff occupancy per sector, time slider enabling the staff manager to preview changes in sectorisation plans on a map. The predicted state of the system in terms of traffic movements and weather condition (e.g., wind) is also displayed and responsive to the time slider.
Step	4	The AI assistant may propose several alternative sector plans, each with a different probability value (based on historical data) and robustness score depending on available ATCO staff, fluctuations in predicted traffic load, and uncertainty in weather forecasts. Using the time slider, the staff manager can evaluate the probability and robustness scores for different times within the maximum look-ahead time horizon.
Step	5	The staff manager interacts with the suggested sector plan in one of the following ways: 1) accept the top-rated AI suggestion and implement it; 2) nudge the AI suggestions by making small changes (e.g., one merge or split); 3) revise large sections of the plan (e.g., revise multiple sectorisation events across various time horizons).
Post-condition	6	The AI assistant monitors changes in predicted system and environmental states. When updated information deviates from the information and data that was used for the implemented sector plan, the AI assistant issues an alert, triggering the staff manager to go back to Step 2.

Operational Scenario		
Type	***No.***	***Description***
Pre-condition	1	The system operates with normal settings.
Step	2	An unexpected airspace reservation is requested by the military.
Step	3	The AI system determines the need for new routes and/or sectorization through real-time monitoring of air traffic flight execution
Step	4	The AI system calculates alternative course deviations considering sector capacity limits".
Step	5	The proposed solution is displayed to the supervisor for review.
Step	6	The supervisor either approves the new route or intervenes to improve upon the AI system advice.
Post-condition	7	The adjusted route plan is executed, and delays are minimized to the greatest extent possible.

5 Instances of the Operational Scenario

Instances	
Name	***Description***
Airspace Reservation	Military Airspace reservation could be for various reasons: Military operations/testing (artillery firing, guided missiles, aerobatics and others). Depending on the FIR, there could be multiple activated military airspaces.

POWER GRID SCENARIOS

Page intentionally left blank.

1 Name of the Operational Scenario

ID	*Application Domain(s)*	*Name of Operational Scenario*
OPSCE-UC1.P-1-001	Power Grid	Remedial actions

2 Version management

Version Management			
Version No.	*Date*	*Name of Author(s)*	*Changes*
0.1	13.01.2025	B. LEMETAYER (RTE)	Creation of the document
1.0	19.02.2025	B. LEMETAYER (RTE)	Final version

3 Narrative of the Operational Scenario

Description of Operational Scenario
Use case
UC1.Power Grid - AI assistant
Complete description
This operational scenario is built based on the scenario "Preventive action to grant N or N-1 situation security in case of unplanned outage" from UC1.Power Grid - AI assistant: • In case of unplanned outage, the AI assistant considers the operational context to provide action to ensure grid security to the human operator, • The AI assistant raises warnings in anticipation to the human operator and provides associated action recommendations. This operational scenario should allow **evaluating technical performance and scalability** of the AI-based assistant, as part of WP4 evaluation objectives.

4 Detailed steps

Operational Scenario		
Type	*No.*	*Description*
Pre-condition	1	The system operates with a normal state in real time
Step	2	An unexpected outage occurs, e.g., a line outage
Step	3	The AI assistant detects the outage through real-time monitoring of grid state, and calculates the state of the environment following the outage
Step	4	In case at least one overload is detected, the AI assistant calculates remedial actions considering the outage and the grid state
Step	5	The proposed remedial action is displayed to the human operator as a recommendation for action: • The AI assistant should always provide several recommendations which represent different trade-offs of different objectives • for each recommendation, the root contingency and the corresponding consequences are detailed with corresponding KPIs If no recommendation can be calculated, the AI assistant raises an explicit alert to the human operator
Step	6	The AI assistant checks that the transmission grid security is ensured until 3h hours ahead: • it considers the possible outage of all lines as a list of contingencies (this list depends on the operational policies of the TSO, based on the EU regulation), • for each contingency, it calculates the state of the environment following the outage
Step	7	In case at least one overload is detected, the AI assistant calculates a remedial action considering the outage and the grid state at the relevant time step. The proposed remedial actions are updated after each timestep, considering the grid state (e.g. if the topology changes)

Operational Scenario		
Type	**No.**	**Description**
Step	8	For each forecast and overload, the proposed remedial actions are displayed to the human operator as a recommendation for action: • there can be up to 3 proposed recommendations, • for each recommendation, the concerned timestep, the root contingency and the corresponding consequences are detailed If no recommendation can be calculated, the AI assistant raises an explicit alert to the human operator
Step	9	The human operator: • Either approves the recommendation, • Or decides to go for another remedial action.
Post-condition	10	The selected remediation action is implemented on the grid (e.g. topology change) and the environment's state is updated
Post-condition	11	Episode data is logged.

5 Instances of the Operational Scenario

Instances	
Name	**Description**
Base environment	Predefined synthetic grid2op environment (use *make([environment name])* in Grid2Op) *Note: it is possible to generate additional synthetic data without limitation*
Real data (RTE)	Real data of RTE grid, to be used in a private way for confidentiality reasons. Structure and topology of the grid could however be disclosed.
Real data (TenneT)	Real data of TenneT grid, to be used in a private way for confidentiality reasons. Structure and topology of the grid could however be disclosed.

1 Name of the Operational Scenario

ID	*Application Domain(s)*	*Name of Operational Scenario*
OPSCE-UC1.P-2-002	Power Grid	AI assistant learns

2 Version management

Version Management			
Version No.	*Date*	*Name of Author(s)*	*Changes*
0.1	13.01.2025	B. LEMETAYER (RTE)	Creation of the document
1.0	19.02.2025	B. LEMETAYER (RTE)	Final version

3 Narrative of the Operational Scenario

Description of Operational Scenario
Use case
UC1.Power Grid - AI assistant
Complete description
This operational scenario is built based on the scenario "AI assistant learns from human operator" from UC1.Power Grid - AI assistant: The AI assistant updates its list of recommendations with actions that were performed by the human operator. It relies on "Remedial actions" (OPSCE-UC1.P-1-001) operational scenario. This operational scenario should allow **evaluating optimal AI / human balance** of the AI-based assistant, as part of WP4 evaluation objectives.

4 Detailed steps

Operational Scenario		
Type	*No.*	*Description*
Pre-condition	1	"Remedial actions" (OPSCE-UC1.P-1-001) operational scenario is run on an environment env_1 with an AI assistant A_1.
Pre-condition	2	Episode data obtained at step 1 are logged.
Pre-condition	3	AI assistant A_1 is fine-tuned on action logged, which gives another AI assistant A_2.
Step	4	"Remedial actions" (OPSCE-UC1.P-1-001) operational scenario is run on another environment env_2 (with different data from environment env_1), using the AI assistant A_1.
Step	5	"Remedial actions" (OPSCE-UC1.P-1-001) operational scenario is run on environment env_2, using the fine-tuned AI assistant A_2.
Post-condition	6	Episode data obtained at step 4 are logged.
Post-condition	7	Episode data obtained at step 5 are logged.
Post-condition	8	The results of steps 6 and 7 are compared by using relevant KPIs

5 Instances of the Operational Scenario

Instances	
Name	*Description*
Base environment	Predefined synthetic grid2op environment (use *make([environment name])* in Grid2Op) *Note: it is possible to generate additional synthetic data without limitation*
Real data (RTE)	Real data of RTE grid, to be used in a private way for confidentiality reasons. Structure and topology of the grid could however be disclosed.
Real data (TenneT)	Real data of TenneT grid, to be used in a private way for confidentiality reasons. Structure and topology of the grid could however be disclosed.

1 Name of the Operational Scenario

ID	Application Domain(s)	Name of Operational Scenario
OPSCE-UC2.P-1-003	Power Grid	Real world conditions

2 Version management

Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	13.01.2025	B. LEMETAYER (RTE)	Creation of the document
0.2	10.02.2025	R. BESSA (INESC TEC)	Details on uncertainty
1.0	19.02.2025	B. LEMETAYER (RTE)	Final version

3 Narrative of the Operational Scenario

Description of Operational Scenario
Use case
UC2.Power Grid – Sim2Real
Complete description
This operational scenario is built based on the scenario "Adaptation to real-world conditions" from UC2.Power Grid – Sim2Real: AI assistant's robustness is tested on bad or low-quality data. It relies on "Remedial actions" (OPSCE-UC1.P-1-001) operational scenario. This operational scenario should allow **evaluating Robustness and safety** of the AI-based assistant, as part of WP4 evaluation objectives.

4 Detailed steps

Operational Scenario		
Type	**No.**	**Description**
Pre-condition	1	"Remedial actions" (OPSCE-UC1.P-1-001) operational scenario is run on an environment env_{base} with an AI assistant A_1.
Pre-condition	2	Episode data obtained at step 1 are logged.
Pre-condition	3	AI assistant's perception of the environment env_{base} is altered with perturbations applied to the input space of the AI system. This is done by applying one of the methodologies defined in Task 4.2 to a given environment's data, among following possibilities: • Perturbation agent, • NoisyObservation object, • etc. This modification is implemented into environment env_{base}, yielding a new environment env_{real}.
Step	4	"Remedial actions" (OPSCE-UC1.P-1-001) operational scenario is run on the environment env_{real}, using AI assistant A_1.
Step	5	At each step of run operational scenarios, when displaying action recommendations to the human operator, the AI assistant makes the human operator aware of the epistemic uncertainty associated to each recommended action, which can be used later by the human to select the action in a multi-objective setting.
Post-condition	6	Episode data obtained at step 4 are logged.
Post-condition	7	The results of steps 2 and 6 are compared by using relevant KPIs.

5 Instances of the Operational Scenario

Instances	
Name	***Description***
Base environment	Predefined synthetic grid2op environment (use *make([environment name])* in Grid2Op) *Note: it is possible to generate additional synthetic data without limitation*
Real data (RTE)	Real data of RTE grid, to be used in a private way for confidentiality reasons. Structure and topology of the grid could however be disclosed.
Real data (TenneT)	Real data of TenneT grid, to be used in a private way for confidentiality reasons. Structure and topology of the grid could however be disclosed.

RAILWAY SCENARIOS

Page intentionally left blank.

1 Name of the Operational Scenario

ID	Application Domain(s)	Name of Operational Scenario
OPSCE-UC1.R-1-004	Railway	Re-Scheduling at the Occurrence of Infrastructure malfunction

2 Version management

Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	07.02.2025	R. LIESSNER (DB)	Creation of the document
1.0	07.03.2025	R. LIESSNER (DB), A. EGLI (SBB), D. BOOS (SBB), T. WAEFLER (FHNW), M. SCHNEIDER (FLATLAND)	Finalization for D4.1

3 Narrative of the Operational Scenario

Description of Operational Scenario
Use case
UC1.Railway – Re-Scheduling at the Occurrence of Infrastructure Malfunction
Complete description
This operational scenario addresses the automated and AI-assisted rescheduling of railway operations when an infrastructure malfunction occurs. Infrastructure malfunctions may include track blockages, switch failures, overhead power failures, or signal breakdowns that disrupt normal train operations. To maintain service continuity and minimize delays, the AI-based Traffic Management System (TMS) continuously monitors infrastructure conditions, detects faults in real-time, and recommends adaptive re-scheduling strategies. Human dispatchers oversee the AI recommendations, validate them, and implement necessary operational changes to ensure optimal railway functionality. Typical remedial actions involve changing the order of trains (re-scheduling). However, other routes can also be selected (re-routing).

4 Detailed steps

Operational Scenario		
Type	**No.**	**Description**
Pre-condition	1	The system operates under normal conditions with no significant infrastructure disruptions.
Step	2	A sudden infrastructure failure (e.g., blocked track, signal failure) is detected by the AI-based monitoring system.
Step	3	The AI system assesses the impact of the malfunction and determines alternative routing and scheduling options.
Step	4	The AI-based Traffic Management System (TMS) generates a revised train schedule to minimize disruption.
Step	5	Human dispatchers take an active role in evaluating AI suggestions and decision-making.
Step	6	The AI system continuously monitors the situation and adjusts operations based on real-time conditions.
Post-condition	7	Train services resume with minimal delays, ensuring operational stability and passenger satisfaction.

5 Instances of the Operational Scenario

Instances	
Name	***Description***
Track Blockage	A fallen tree or debris obstructs the track, requiring re-routing and delay minimization.
Signal Failure	A major signal malfunction prevents normal operations, requiring AI-guided manual intervention.
Overhead Power Failure	A power supply disruption affects electrified train services, requiring re-scheduling and alternative train operations.

1 Name of the Operational Scenario

ID	Application Domain(s)	Name of Operational Scenario
OPSCE-UC1.R-2-005	Railway	Emergency Response to Adverse Weather conditions

2 Version management

Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	07.02.2025	R. LIESSNER (DB)	Creation of the document
1.0	07.03.2025	R. LIESSNER (DB), A. EGLI (SBB), D. BOOS (SBB), T. WAEFLER (FHNW), M. SCHNEIDER (FLATLAND)	Finalization for D4.1

3 Narrative of the Operational Scenario

Description of Operational Scenario
Use case
UC1.Railway – Emergency Response to Weather Challenges
Complete description
This operational scenario addresses emergency responses required when severe weather events impact railway operations. The scenario considers various weather conditions, such as heavy snowfall, flooding, strong winds, and storms, which may cause disruptions such as blocked tracks, overhead power failures, landslides, or reduced visibility for train operators. To ensure continued service delivery and passenger safety, the AI-based Traffic Management System (TMS) and human dispatchers work in coordination to re-schedule train operations dynamically, adjust speed limits, reroute trains, and optimize available infrastructure to minimize delays and avoid potential hazards.

4 Detailed steps

Operational Scenario		
Type	**No.**	**Description**
Pre-condition	1	The system operates under normal conditions with no significant weather-related disruptions.
Step	2	A severe weather event disrupted the power line between two stations. The track had to be closed and the closure is detected by the AI-system.
Step	3	The AI system assesses the weather impact and identifies affected areas within the railway network.
Step	4	The AI-based Traffic Management System (TMS) generates alternative schedules and reroutes to ensure minimal disruptions.
Step	5	Human dispatchers take an active role in evaluating AI suggestions and decision-making.
Step	6	The system continuously monitors operations, adjusting to new weather impacts on infrastructure availability if required.
Post-condition	7	Train operations continue with minimal disruptions, ensuring safety and service continuity.

5 Instances of the Operational Scenario

Instances	
Name	**Description**
Heavy Snowfall	A snowstorm affects key railway corridors, causing track obstructions and delays.

Instances	
Name	***Description***
Flooding	Heavy rain leads to waterlogged tracks, requiring rerouting and speed adjustments.
Storm Disruptions	Strong winds impact catenary wires and signal visibility, requiring schedule adjustments.

1 Name of the Operational Scenario

ID	Application Domain(s)	Name of Operational Scenario
OPSCE-UC1.R-3-006	Railway	Partial closure of a large station

2 Version management

Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	20.01.2025	R. LIESSNER (DB)	Creation of the document
1.0	07.03.2025	R. LIESSNER (DB), A. EGLI (SBB), D. BOOS (SBB), T. WAEFLER (FHNW), M. SCHNEIDER (FLATLAND)	Finalization for D4.1

3 Narrative of the Operational Scenario

Description of Operational Scenario
Use case
UC1.Railway - Closure of a large station or parts thereof
Complete description
This operational scenario addresses the challenge of adjusting railway operations when a track, e.g. in a station, is suddenly closed due to unforeseen emergencies, or security threats. The underlying goal is to introduce a major disruption that has a large and long-term impact on events: the partial closure of a major station (one or multiple tracks) significantly disrupts train schedules, passenger flow, and freight logistics, requiring an immediate and effective re-scheduling response. The AI-based Traffic Management System (TMS) and human operators collaborate to generate alternative routing plans and communicate changes efficiently: the AI assistant should solve this as independently as possible. The human is only there for active monitoring. The objective is to ensure minimal disruption to railway services while maintaining operational safety and efficiency.

4 Detailed steps

Operational Scenario		
Type	**No.**	**Description**
Pre-condition	1	The railway network operates under normal conditions with scheduled services running through the station.
Step	2	An emergency or planned event leads to the partial closure of a large station.
Step	3	The AI-based Traffic Management System (TMS) detects the closure and initiates an assessment of network impact.
Step	4	The AI-based Traffic Management System (TMS) generates a revised train schedule to minimize disruption.
Step	5	Human dispatchers take an active role in evaluating AI suggestions and decision-making.
Step	6	Continuous monitoring ensures adjustments can be made based on real-time operational data.
Post-condition	7	Train operations continue smoothly, ensuring minimal disruptions and passenger inconvenience.

5 Instances of the Operational Scenario

Instances	
Name	***Description***
Emergency Closure	A security threat forces the immediate partial shutdown of a major station, requiring instant rescheduling.
Infrastructure Failure	Structural damage or maintenance work leads to an unplanned long-term closure.
Passenger Overload	The closure of a station results in significant congestion at adjacent stations.

1 Name of the Operational Scenario

ID	Application Domain(s)	Name of Operational Scenario
OPSCE-UC2.R-1-007	Railway	Reactive Re-Scheduling

2 Version management

Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	07.02.2025	R. LIESSNER (DB)	Creation of the document
1.0	07.03.2025	R. LIESSNER (DB), A. EGLI (SBB), D. BOOS (SBB), T. WAEFLER (FHNW), M. SCHNEIDER (FLATLAND)	Finalization for D4.1

3 Narrative of the Operational Scenario

Description of Operational Scenario
Use case
UC2.Railway – Reactive Re-Scheduling
Complete description
Reactive Re-Scheduling refers to a human-AI collaborative process of adapting train schedules after an unexpected disruption has already occurred. Such disruptions may include infrastructure failures, train breakdowns, extreme weather conditions, or sudden passenger demand fluctuations. Both, the AI and the human continuously monitor the network in order to predict potential disruptions.by monitoring monitors railway operations. The AI-based Traffic Management System (TMS) provides optimized re-scheduling solutions. Human dispatchers review AI-generated suggestions, validate changes, and implement new schedules to minimize service disruption and improve operational resilience.

4 Detailed steps

Operational Scenario		
Type	**No.**	**Description**
Pre-condition	1	The railway network operates under normal conditions with an optimized pre-planned schedule.
Step	2	A sudden disruption (e.g., track failure, train delay) is detected by the AI-based monitoring system and/or by the human dispatcher..
Step	3	The AI system assesses the impact of the disruption on the network and identifies affected train routes.
Step	4	The AI-based Traffic Management System (TMS) generates optimized re-scheduling recommendations.
Step	5	Human dispatchers review AI-generated schedules, validate or modify changes as necessary.
Step	6	The AI system continuously monitors the adjusted schedule and makes further refinements if needed.
Post-condition	7	Train operations continue with minimal disruptions, ensuring service reliability.

5 Instances of the Operational Scenario

Instances	
Name	**Description**
Train Breakdown	A mechanical failure causes a train to stop unexpectedly, requiring schedule modifications.

Instances	
Name	***Description***
Signal Malfunction	A failure in the signaling system leads to train delays, requiring immediate re-scheduling.
Sudden Passenger Demand	Unanticipated passenger surges require reallocation of train resources and modified timetables.

1 Name of the Operational Scenario

ID	Application Domain(s)	Name of Operational Scenario
OPSCE-UC2.R-2-008	Railway	Co-learning for Reactive Re-Scheduling

2 Version management

Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	07.02.2025	R. LIESSNER (DB)	Creation of the document
1.0	07.03.2025	R. LIESSNER (DB), A. EGLI (SBB), D. BOOS (SBB), T. WAEFLER (FHNW), M. SCHNEIDER (FLATLAND)	Finalization for D4.1

3 Narrative of the Operational Scenario

Description of Operational Scenario
Use case
UC2.Railway – Co-learning for Reactive Re-Scheduling
Complete description
Co-learning for Reactive Re-Scheduling refers to the iterative learning process between AI systems and human dispatchers in responding to unexpected railway disruptions. This scenario focuses on refining the collaboration between human decision-makers and AI-based Traffic Management Systems (TMS) by integrating feedback loops that enhance both AI learning models and human expertise over time. During reactive re-scheduling, AI continuously analyzes real-time data, detects deviations, and suggests optimized rescheduling options. Human dispatchers provide contextual insights and validate AI decisions, enabling mutual adaptation and improved system performance. The goal is to optimize the decision-making process and enhance the effectiveness of both AI-generated solutions and human responses. Human dispatchers are supported by AI to continuously improve their expertise developing response strategies collaboratively with the AI.

4 Detailed steps

Operational Scenario		
Type	**No.**	**Description**
Pre-condition	1	The railway network operates under normal conditions with an optimized pre-planned schedule.
Step	2	A sudden disruption (e.g., track failure, train delay) is detected by the AI-based monitoring system an/or by the human dispatcher.
Step	3	The AI system assesses the impact of the disruption and generates an initial set of rescheduling recommendations.
Step	4	Human dispatchers review AI-generated suggestions, provide feedback, and refine recommendations. The human dispatcher selects the recommendation to be executed.
Step	5	AI system adjusts future recommendations based on human feedback.
Step	6	Human-AI collaborative continuous monitoring is maintained to improve response strategies.
Post-condition	7	The AI system enhances its learning models while AI explicitly supports human dispatchers in improve their decision-making expertise.

5 Instances of the Operational Scenario

Instances	
Name	***Description***
AI-Dispatcher Collaboration	Dispatchers override AI suggestions and their decisions are used to refine AI learning.
Pattern Recognition	AI identifies recurring disruption patterns and adapts re-scheduling strategies accordingly.
Decision Alignment	AI and human decisions converge over time, leading to more efficient re-scheduling outcomes.
Human learning	Human learning process regarding their decision-making capabilities are explicitly supported by AI.

1 Name of the Operational Scenario

ID	Application Domain(s)	Name of Operational Scenario
OPSCE-UC2.R-3-009	Railway	Proactive Re-Scheduling

2 Version management

Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	07.02.2025	R. LIESSNER (DB)	Creation of the document
1.0	07.03.2025	R. LIESSNER (DB), A. EGLI (SBB), D. BOOS (SBB), T. WAEFLER (FHNW), M. SCHNEIDER (FLATLAND)	Finalization for D4.1

3 Narrative of the Operational Scenario

Description of Operational Scenario
Use case
UC2.Railway – Proactive Re-Scheduling
Complete description
Proactive Re-Scheduling refers to a human-AI collaborative to predicting and mitigating potential disruptions before they occur. Unlike reactive re-scheduling, this scenario focuses on identifying early warning signals, analyzing patterns, and proactively adjusting train schedules to minimize operational disruptions. Both, the AI and the human continuously monitor the network in order to predict potential disruptions. The AI-based support system for Traffic Management suggests optimized scheduling adjustments. Human dispatchers validate these recommendations, ensuring operational feasibility and safety compliance while improving response readiness for unforeseen circumstances.

4 Detailed steps

Operational Scenario		
Type	**No.**	**Description**
Pre-condition	1	The railway network operates under normal conditions with an optimized pre-planned schedule.
Step	2	The AI system continuously monitors railway operations and collects data on potential weak signals, e.g. trains that are slightly delayed and will lead to future conflicts. The human dispatchers monitors the network for weak signals that indicate potential disruptions.
Step	3	The AI system and/or the human dispatchers detect early warning signs of potential disruptions.
Step	4	AI-based Traffic Management System (TMS) generates alternative scheduling adjustments.
Step	5	Human dispatchers review AI-generated recommendations and validate or modify proposed changes.
Step	6	The updated schedule is implemented, adjusting train routes, speeds, and platform assignments as needed.
Step	7	Continuous monitoring ensures additional refinements based on real-time railway conditions.
Post-condition	8	Operational resilience is improved, and potential delays are minimized before disruptions occur.

5 Instances of the Operational Scenario

Instances	
Name	***Description***
Weather-Based Adjustments	AI predicts potential weather-related disruptions and adjusts train schedules proactively.
Passenger Flow Optimization	AI analyzes historical passenger trends to optimize peak-hour scheduling.
Human expertise	Based on their experience and tacit knowledge human dispatcher recognize patterns in railway conditions and hence potential disturbances.

1 Name of the Operational Scenario

ID	Application Domain(s)	Name of Operational Scenario
OPSCE-UC2.R-4-010	Railway	Co-learning for Proactive Re-Scheduling

2 Version management

Version Management			
Version No.	**Date**	**Name of Author(s)**	**Changes**
0.1	07.02.2025	R. LIESSNER (DB)	Creation of the document
1.0	07.03.2025	R. LIESSNER (DB), A. EGLI (SBB), D. BOOS (SBB), T. WAEFLER (FHNW), M. SCHNEIDER (FLATLAND)	Finalization for D4.1

3 Narrative of the Operational Scenario

Description of Operational Scenario
Use case
UC2.Railway – Co-learning for Proactive Re-Scheduling
Complete description
Co-learning for Proactive Re-Scheduling focuses on the mutual learning process between AI systems and human dispatchers when implementing proactive re-scheduling strategies. The goal is to improve system accuracy, operational efficiency, and human expertise by integrating human feedback into AI learning models while enabling human dispatchers to better understand AI-driven scheduling suggestions. Both, the AI and the human continuously monitor the network in order to predict potential disruptions. The AI-based Traffic Management System (TMS) proposes scheduling adjustments. Human dispatchers validate and modify these recommendations based on experience and operational context. AI then incorporates dispatcher feedback to refine its predictive models, ensuring a continuous improvement cycle. Human dispatchers are supported by AI to continuously improve their expertise in detecting potential disturbances and to develop response strategies collaboratively with the AI.

4 Detailed steps

Operational Scenario		
Type	**No.**	**Description**
Pre-condition	1	The railway network operates under normal conditions with an optimized pre-planned schedule.
Step	2	The AI system continuously collects and analyzes data for potential disruptions. The human dispatchers monitor the network for weak signals that indicate potential disruptions.
Step	3	AI and/or the human dispatchers detect early signs of potential delays or conflicts. AI generates preliminary schedule adjustments.
Step	4	The AI-based Traffic Management System (TMS) presents scheduling recommendations to human dispatchers.
Step	5	Human dispatchers review AI-generated recommendations and provide contextual feedback.
Step	6	AI system adjusts recommendations based on human feedback.
Step	7	The human dispatcher takes responsibility for the final implementation, meaning they input the instructions for rerouting and/or time adjustments (sequence decisions), etc., and the machine transmits the information to the "simulation" so that the new schedule can be executed.

Operational Scenario		
Type	**No.**	**Description**
Step	8	Continuous monitoring ensures adaptive learning cycles, enhancing the collaboration between AI and human decision-makers.
Post-condition	9	AI and human dispatchers improve in detecting potential disturbances and develop improved response strategies, increasing railway resilience and efficiency.

5 Instances of the Operational Scenario

Instances	
Name	**Description**
AI-Human Collaboration	Human dispatchers validate AI recommendations and AI integrates feedback for future improvements. AI supports improving corresponding human expertise.
Predictive Model Refinement	AI identifies trends and adjusts its predictive scheduling models based on dispatcher feedback.
Human predictive monitoring expertise.	AI supports humans' continuous learning regarding the detection of weak signals that indicate potential disturbances.
Optimized Decision-Making	Human-AI synergy results in progressively improved proactive scheduling decisions.

ANNEX 3 – RISK ASSESSMENT

QUESTIONNAIRE

OBJECTIVE

This document is intended to guide the identification and analysis of risks in the use cases of AI4REALNET. In the scope of Task 4.2 and aligned with Article 15 ("Accuracy, robustness, cybersecurity") of the AI Act, **the goal is to identify the risks associated with errors, faults, or inconsistencies that may occur within the system or the environment in which the system operates, due to their interaction with natural persons or other systems.**

This process is based on the definitions and processes included in the **ISO/IEC 31000:2018** and **ISO/IEC 23894:2023** standards to align the risk assessment methodology in AI4REALNET with the current European standards. This approach lays the ground for the regulatory assessment planned in Task 4.4, which will verify the compliance of AI4REALNET solutions with the current EU regulations and requirements for high-risk applications.

The formalization of risks follows the multi-component framework proposed in the AI4REALNET conceptual framework (D1.1), which is based on consolidated methodologies in disaster risk reduction – e.g., the IPCC risk framework[15] – and allows to cover the caveats of the EU AI Act, which lack representation of the risk sources (hazards) and factors that make involved actors vulnerable to these hazards. Including these points in the assessment allows a more precise understanding of the effectiveness of mitigation strategies later. In the later stage of robustness and safety assessment, separating risk into risk components allows a more accurate comparison of different risk sources and quantification of the severity of their outcomes for the system and stakeholders.

SCOPE

The framework is designed to offer a standardised assessment of both technical and socio-ethical risks for technical components, human actors and material resources. However, **Task 4.2 concentrates on the technical risks that can influence the AI system**. Hence, **only risks that impact the robustness and safety of the AI system should be considered for this part of the project**.

Consider only the **use cases from WP1** and the fact that the AI model can be **reinforcement learning and/or supervised learning** mainly based on neural networks. The analysis should consider **a full operational scenario of the AI system** (although the project will only reach a proof-of-concept stage). Hence, this process should also analyse risks arising when the system runs in a real-world environment.

We recommend that the risk analysis be based on the technical requirements derived from AI4REALNET Use Cases in Task 1.4 and the multi-component risk framework from the conceptual

[15] https://www.ipcc.ch/site/assets/uploads/2021/02/Risk-guidance-FINAL_15Feb2021.pdf

framework (D1.1). The assessed risk should be seen as a violation of one of the requirements and, in this way, connected to the previous analysis of AI applications planned for implementation in the project.

Consider the presence of non-AI components, e.g., human operator (which can override AI), (cyber)security measures in place, and mitigation actions in case of security breaches. For instance, in case of an intentional cyber-attack on the AI model, the AI models are generally stored and operated in highly cyber-secure Information Technology (IT) systems. Moreover, in the event of an attack, the previously trained model could be quickly restored.

TERMINOLOGY

Risk – a potential unexpected "effect of uncertainty on objectives"[16]. This corresponds to all unwanted effects on the objectives, that create threats or opportunities.

Example: Deadlock of two trains on a single track. Blackout in the power grid.

Risk source – an event that leads to the risk.

Example: The wrong decision taken by the AI system.

Asset – tangible or intangible resources related to the design and use of the AI that are impacted by the risk. Can relate to individuals, elements in the organization, society, or environment that are being impacted by the risk[17].

"Examples:

Assets and their value to the organization:

— Tangible assets can include data, models, and the AI system itself.

— Intangible assets can include reputation and trust.

Assets of and their value to individuals:

— Tangible assets can include an individual's personal data,

— Intangible assets can include the privacy, health, and safety of an individual.

Assets of and their value to communities and societies:

— Tangible assets can include the environment,

— Intangible assets are likely more values-based, such as socio-cultural beliefs, community knowledge, educational access, and equity."

Hazard – the characterization of the risk source, which involves the estimation of:

- **Rate or probability of occurrence** estimates the frequency of the occurrence of the unwanted event or the probability of such an event happening in a given time interval T.

[16] ISO/IEC 31000:2018, "Risk management Guidelines"

[17] ISO/IEC 23894:2023, "Information technology - Artificial intelligence - Guidance on risk management"

- **Magnitude** Severity of the impact that this hazard can produce in the worst case on a single given asset (e.g., one person, one AI component).

Exposure – the inventory of assets within the range of a specific hazard (e.g., human operator, end users, or other system components). Evaluating the exposure involves quantifying the **number of these assets exposed to the hazard**.

Vulnerability – the propensity of the asset to be impacted by the hazard. It is the result of combined diversity drivers. Drivers may increase the vulnerability (e.g., "being part of populations at risk") or decrease it ("existing safety measures").

Mitigation – refers to solutions to address AI-specific vulnerabilities, such as measures to prevent, detect, respond to, resolve, and control risk sources. Technical solutions to address AI-specific vulnerabilities shall include, where appropriate, such mitigation measures for attacks trying to manipulate the training data set (data poisoning), or pre-trained components used in training (model poisoning), inputs designed to cause the AI model to make a mistake (adversarial examples or model evasion), confidentiality attacks or model flaws.[18]

RISK ASSESSMENT

The proposed procedure for identifying and analysing the risks follows the risk management processes described in the ISO 31000 standard. The risk assessment process comprises the following steps:

- Risk identification.
- Risk analysis.
- Risk evaluation.

Risk evaluation is out of scope for Task 4.2. The steps relevant to Task 4.2 are described in detail below.

RISK IDENTIFICATION

The analysis of the risks (assessment of the risk components) should start with identifying the context of the unwanted situation. The purpose is to customize the risk management process and facilitate understanding and characterizing the risk and its appropriate treatment.

At the core of a risk assessment lies the identification of risks, i.e., the events that cause the unwanted effects. When describing this risk, it is necessary to identify which assets may be affected. One event can lead to different outcomes for different kinds of assets, e.g., too high a current directed to a part of the power grid leading to a congestion problem that can impact end-users, whose households are left without electricity, and the infrastructure if the power grid is damaged. These two different impacts are covered by enumerating two different risks.

The minimum risk description necessary for the risk assessment consists of:

- Short descriptor of the risk (risk name).
- Provide a description of the risk source and the process of how it is being evoked.
- The type of asset that is harmed.

[18] https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf

An example of a risk description is an airplane collision (descriptor). A crash of two airplanes (assets) occurred because of the AI system's wrong decision, which the supervisor (risk source) accepted.

RISK ANALYSIS

During the analysis of each identified risk, we need to assess its specific components:

- **Hazard**, which involves the estimation of:
  - **Rate or probability of occurrence**.
  - **Magnitude**
- **Exposure**.
- **Vulnerability**.

To characterize the hazard, we analyse the risk source by estimating its magnitude and rate/probability. The exposure and vulnerability are defined by assets: the state of the environment outside of the AI system and properties of the assets. These components should be assessed using a clearly documented scale. For example, for assessing the hazard rate, we can choose the scale with the following categories:

very low - once in 10000 decisions

low – once in 1000-10000 decisions

moderate – once in 100-1000 decisions

high – once in 10-100 decisions

very high – once in 10 decisions.

It is worth mentioning that a given risk can have multiple sources and impact various assets. To use the analysis effectively for the risk mitigation strategy, such cases must be described in detail for each hazard or asset, as they may have different risk components. The full quantification of risk demands the assessment of the risk value for each source/asset.

EXAMPLE OF THE RISK ASSESSMENT

Table 14 provides examples of risk identification and analysis according to the described procedure. It includes three risks: two risks that impact multiple assets (intentional/unintentional attacks and deadlock of two trains) and a risk that can result from two risk sources (The AI system's decisions systematically favour certain train lines, which can happen due to misclassification or an attack, that changes the behaviour of the model). These examples showcase how the separation of risk components according to the context can illustrate the differences in various risk situations.

The first example is a detailed analysis of the consequences an external perturbation can have on an AI system. In this case, the asset is defined as an AI system, and the risk sources are distinguished depending on the targeted part of the system, providing a more detailed overview of the impact depending on the target of the attack. This level of detail is required for the later development of risk mitigation strategies, which can be implemented in different parts of the AI system in the case of technical solutions (as mentioned above and in EU AI Act Article 15). **Evaluating the robustness of**

these risks is the main goal of Task 4.2. Therefore, this first example should be considered in all domains.

For the second example of unfair treatment of certain lines, we compare two possible sources of this risk: misclassification due to the performance of the system itself (AI component and human operator combined) and misclassification as an effect of a malicious attack. In the first case, the rate of the hazard is equal to the rate of misclassification, which can be calculated, for example, from the system's accuracy and the likelihood of misclassifications being corrected by the human operator. We are aware that such a situation would also have an influence on more than one type of asset; however, for the sake of understandability, we only focus on one: the passengers. To address the vulnerability, we must define the outcome of the misclassification for a user. The exposure is equal to the number of trains affected by this hazard. For the second case, the rate of the hazard changes to the rate of an attack, and the magnitude is characterized by the likelihood of a change in decision due to the attack. With the change of the risk source, these components also change. However, there is no difference in vulnerability and exposure, as the output does not differ in the two cases: the users are affected independently of the cause. This comparison shows that different risks can lead to the same outcomes, yet because of different sources, different mitigation strategies are needed to prevent their occurrence.

In the third example of deadlock, the hazard depends on the risk source and stays the same for each of the affected assets: passengers, transported goods, and infrastructure. The vulnerability and exposure, however, change depending on the asset being considered.

Risk	Risk source	Asset	Hazard rate (Very Low /Low / Medium / High / Very High)	Hazard magnitude	Vulnerability	Exposure
An intentional (attack) or unintentional external perturbation damages the system's robustness.	Perturbation (or attack) at AI model	AI system[19]	*How frequent these attacks/perturbations could be?*	*Indicate and classify potential impacts. Indicate the range of the potential impact.*	*Are there special features that make the AI model more vulnerable to this risk?* *How quickly we can restore a model in the event of an attack?* *Does an attacker have all the information to define an attack to the model output?*	*How many elements/components of the AI system are exposed to the risk?*
	Perturbation (or attack) at reward/loss function	AI system	*How frequent these attacks/perturbations could be?*	*Indicate and classify potential impacts. Indicate the range of the potential impact.*	*Are there special features that make the AI model more vulnerable to this risk?* *How quickly we can restore a model in the event of an attack?* *Does an attacker have all the information to define an attack to the model output?*	*How many elements/components of the AI system are exposed to the risk?*
	Perturbation (or attack) at action/output space	AI system	*How frequent these attacks/perturbations could be? State the time interval T for the definition.*	*Indicate and classify potential impacts. Indicate the range of the potential impact.*	*Are there special features that make the AI model more vulnerable to this risk?* *How quickly we can restore a model in the event of an attack?* *Does an attacker have all the information to define an attack to the model output?*	*How many elements/components of the AI system are exposed to the risk?*

[19] The risk linked to this source can be analysed for the asset being the entire decision-making system. This may lead to the different assessments of risk components (due to safety measures or robustness properties brought by the non-AI components of the system).

Risk	Risk source	Asset	Hazard rate (Very Low /Low / Medium / High / Very High)	Hazard magnitude	Vulnerability	Exposure
	Perturbation (or attack) at state/input space	AI system	*How frequent these attacks/perturbations could be? State the time interval T for the definition.*	*Indicate and classify potential impacts. Indicate the range of the potential impact.*	*Are there special features that make the AI model more vulnerable to this risk?* *How quickly we can restore a model in the event of an attack?* *Does an attacker have all the information to define an attack to the model output?*	*How many elements/components of the AI system are exposed to the risk?*
The AI system's decisions systematically favour certain lines over others (regional over national)	Misclassification	The passengers of those lines.	Misclassification rate (ex. derived from system's accuracy)	*How likely is it that the misclassification leads to favouring one line over another?*	*What would this unfair treatment mean to the passengers? Can they use other lines? Do their connections mostly include transfer to other trains and hence longer delay time?*	*Is the number of affected passengers low / medium / high?*
	Malicious attack leading to misclassification	The passengers of those lines.	Attack rate	*How likely is it, that the attack would change the decision? the attack efficacy*	*Same*	*Same*
Deadlock of two trains on rail tracks leading to blockage of the line	AI system directed two trains to one track from opposite directions	Passengers	Can be calculated from system's accuracy metrics or KPIs of the use case	*How big are the caused delays? How fast can the system recover?*	*What type of train is it, how big is the capacity?*	*Is the number of passengers of delayed trains low / medium / high?*
	AI system directed two trains to one track from opposite directions	Transported goods	Can be calculated from system's accuracy metrics or KPIs of the use case	*How big are the caused delays? How fast can the system recover?*	*Type of goods: how time sensitive is the delivery, can something be rotten, are there*	*Is the amount of goods being transported low / medium / high?*

Risk	Risk source	Asset	Hazard rate (Very Low /Low / Medium / High / Very High)	Hazard magnitude	Vulnerability	Exposure
					legal penalties for delayed delivery?	
	AI system directed two trains to one track from opposite directions	Infrastructure	Can be calculated from system's accuracy metrics or KPIs of the use case	*How big are the caused delays? How fast can the system recover?*	*Can other trains run on the parallel tracks? How far away is the next lock, that can release the deadlock?*	*Is the number of kilometres of the tracks or number of affected tracks low / medium / high?*

TABLE 14 – RISK ANALYSIS MATRIX

RISK ASSESSMENT RESULTS

AIR TRAFFIC MANAGEMENT

Risk	Risk source	Asset	Rate	Magnitude	Vulnerability	Exposure
An intentional (attack) or unintentional external perturbation damages the system's robustness.	Perturbation (or attack) at AI model	AI system	*How frequent these attacks/perturbations could be?* **Rate: Low** Attacks or perturbations at the AI model level are relatively rare due to the robust cybersecurity measures in place within the ATM industry's IT infrastructure. While possible, such incidents are infrequent.	*Indicate and classify potential impacts. Indicate the range of the potential impact.* An attack or perturbation affecting the AI model would not lead to significant operational disruptions. Potential impacts may include initial incorrect routing or sectorization decisions, but since the AI advisory systems works on a pre-tactical time horizon, tactical operations are not affected.	*Are there special features that make the AI model more vulnerable to this risk?* *How quickly we can restore a model in the event of an attack?* *Does an attacker have all the information to define an attack to the model output?* The AI model is safeguarded within a highly secure IT environment equipped with advanced cybersecurity protocols. Rapid restoration procedures are in place to quickly recover the model in the event of an attack. Attackers are unlikely to have sufficient access or information to effectively compromise the model output. Therefore, the vulnerability of the AI model to such risks is **low**.	*How many elements/components of the AI system are exposed to the risk?* A compromise of the model could affect multiple aspects of ATM operations. Therefore, the exposure is considered medium, as the impact can propagate to several ATM components despite the limited number of elements directly at risk.
	Perturbation (or attack) at reward/loss function	AI system	Attacks or perturbations targeting the reward or loss function are relatively rare due to robust security measures in place. Such functions are typically internal components of the AI system and are not easily accessible. However, insider threats or sophisticated external attacks could potentially lead to such incidents.	A successful attack or perturbation on the reward/loss function could significantly degrade the AI system's performance. Potential impacts include: **Incorrect Decision-Making:** The AI could learn inappropriate policies, leading to suboptimal or unsafe operational decisions. **Safety Risks:** Misguided actions could compromise passenger safety or lead to accidents. **Operational Disruptions:** Scheduling errors, delays, or resource misallocations could occur. The range of impact could span from minor operational disturbances to severe safety incidents.	**Special Features Affecting Vulnerability:** The AI model's vulnerability depends on the security of the training environment and the integrity checks in place for the reward/loss functions. If these components are well-protected and monitored, vulnerability is low. **Restoration Time:** The model can be restored relatively quickly if backups and version control systems are in place. Retraining may be necessary, which could take from hours to days, depending on the complexity. Are there special features that make the AI model more vulnerable to this risk? How quickly we can restore a model in the event of an attack? Does an attacker have all the information to define an attack to the model output? **Attacker's Information Access:** It is unlikely that an external attacker would have sufficient access to manipulate the reward/loss function without insider assistance. Therefore, the risk is mostly from internal threats or advanced persistent threats that have breached security layers.	*How many elements/components of the AI system are exposed to the risk?* The primary component exposed is the AI system's training module where the reward/loss functions are defined and utilized. While this is a singular component, its central role means that a successful attack could affect the entire AI system's behavior. Therefore, the exposure is considered **medium**: **Direct Exposure:** The reward/loss function within the AI system. **Indirect Impact:** The compromised function can influence all decision-making processes of the AI, thereby affecting multiple system components and operational areas.

Risk	Risk source	Asset	Rate	Magnitude	Vulnerability	Exposure
	Perturbation (or attack) at action/output space	AI system	*How frequent these attacks/perturbations could be? State the time interval T for the definition.* **Rate: Moderate** Attacks or perturbations at the action/output space could occur with a moderate frequency. Since the outputs of the AI system will be submitted to human operators, they may be more accessible to attackers compared to internal components like the model or reward function. If the communication channels are not fully secured, the risk of such perturbations increases.	*Indicate and classify potential impacts. Indicate the range of the potential impact.* A successful attack or perturbation on the action/output space can have significant impacts: **Operational Disruptions:** Incorrect outputs could lead to ATCO scheduling conflicts and incorrect sectorizations. **Safety Risks:** Implementing erroneous sectorization decisions could compromise passenger safety, potentially leading to accidents or hazardous situations. The range of potential impacts spans from minor inconveniences to severe safety incidents affecting multiple stakeholders.	*Are there special features that make the AI model more vulnerable to this risk?* *How quickly we can restore a model in the event of an attack?* *Does an attacker have all the information to define an attack to the model output?* **Special Features Affecting Vulnerability:** The AI model may be more vulnerable if output data is transmitted over unsecured networks or if there is a lack of robust authentication and validation mechanisms for the outputs. **Restoration Time:** Since the attack targets the outputs rather than the model itself, restoration involves identifying the compromised outputs and securing the communication channels. With effective monitoring and incident response protocols, this can be accomplished relatively quickly—potentially within hours. **Attacker's Information Access:** An attacker does not necessarily need in-depth knowledge of the AI model to manipulate the outputs. If they can intercept or alter the data in transit, they can influence the system's actions. Therefore, the risk is higher if outputs are accessible without strong security measures.	*How many elements/components of the AI system are exposed to the risk?* The exposure is considered **High** because multiple components depend on the AI system's outputs: **Direct Exposure:** All communication interfaces where the AI outputs are transmitted, including data links to operational systems and human-machine interfaces. **Indirect Impact:** Downstream systems and processes that act on the AI's outputs, such as alarm systems and operational staff, could be affected by compromised outputs. A significant portion of the AI system and its connected components are exposed to this risk, which can propagate through the operational network and impact overall ATM safety and efficiency performance.

Risk	Risk source	Asset	Rate	Magnitude	Vulnerability	Exposure
	Perturbation (or attack) at state/input space	AI system	*How frequent these attacks/perturbations could be? State the time interval T for the definition.* **Rate**: **Moderate to High** Attacks or perturbations at the state/input space can occur with a moderate to high frequency. Since the AI system relies on a multitude of input data sources—including surveillance systems, external databases, communication networks, and user inputs—there is a risk that these inputs can be intentionally manipulated or inadvertently corrupted. If input channels are not fully secured or lack proper validation mechanisms, the likelihood of such perturbations increases. For example, these attacks or perturbations might occur once in every X to input instances, depending on the system's exposure and the robustness of security measures in place.	*Indicate and classify potential impacts. Indicate the range of the potential impact.* A successful perturbation or attack on the state/input space can have significant impacts: **Operational Disruptions:** Incorrect or manipulated input data can lead the AI system to make erroneous decisions, resulting in scheduling conflicts, and sector misallocations. The range of potential impacts spans from minor operational inconveniences to severe safety incidents affecting multiple stakeholders, including passengers, staff, and infrastructure.	*Are there special features that make the AI model more vulnerable to this risk?* *How quickly we can restore a model in the event of an attack?* *Does an attacker have all the information to define an attack to the model output?* **Special Features Affecting Vulnerability:** **Lack of Input Validation:** If the AI model does not have robust input validation or anomaly detection mechanisms, it is more susceptible to accepting and acting upon corrupted or malicious data. **Dependency on External Data Sources:** Heavy reliance on external data without adequate verification increases vulnerability. **Unsecured Sensors and Communication Channels:** Use of unencrypted data transmission and unsecured sensor networks makes it easier for attackers to introduce perturbations. **Restoration Time:** - Since the attack targets the input data rather than the AI model itself, restoration focuses on identifying and filtering out corrupted inputs and securing the data channels. - With effective monitoring systems and rapid incident response protocols, the issue can be mitigated relatively quickly—potentially within minutes to a few hours. - However, if the perturbation has caused cascading effects or if the source of the attack is not immediately identifiable, full restoration could take longer. **Attacker's Information Access:** - An attacker may not require extensive knowledge of the AI model to manipulate the inputs. Access to input interfaces, sensors, or data transmission paths may suffice. - If input sources are accessible without strong authentication and encryption, the risk of successful attacks increases. - Therefore, the vulnerability is moderate to high, depending on the existing security measures protecting the input channels.	*How many elements/components of the AI system are exposed to the risk?* The exposure is considered **High** because multiple components depend on accurate and reliable input data: **Direct Exposure:** o All input interfaces are exposed, including: **§ Sensors:** Surveillance systems measuring aircraft states. **§ External Data Feeds:** Information from weather services and Central Flow Managements Units **§ User Inputs:** Data entered by human users (FMP supervisor). **§ Communication Networks:** Channels over which input data is transmitted to the AI system. **Indirect Impact:** o Components and processes relying on the AI system's decisions are indirectly exposed, such as: **§ Operational Control Systems:** Systems that manage aircraft scheduling and routing. **§ Safety Mechanisms:** Conflict detection systems (e.g., Medium to Long Term Conflict Warning). **§ Staff:** Human actors who depend on the AI system for safe and efficient operations. **Overall Exposure:** o Given that the AI system's functionality is heavily dependent on input data, a significant portion of the system and its connected components are exposed to this risk. o The impact can propagate throughout the operational network, potentially affecting the entire ATM system's functionality and safety. o Therefore, the exposure level necessitates robust security measures across all input channels to mitigate the risk effectively.

Risk	Risk source	Asset	Rate	Magnitude	Vulnerability	Exposure
The AI system's decisions systematically favour certain sectorizations and routings over others	Misclassification	The airlines and passengers of those routings.	Misclassification rate (ex. derived from system's accuracy) **Misclassification Rate:** Derived from the system's accuracy. For instance, if the AI system has a 95% accuracy, the misclassification rate is 5%. **Likelihood of Favoritism Due to Misclassification:** Not all misclassifications result in favoring certain sectorizations and routings. Suppose **10%** of misclassifications lead to favoring one sector and/or routing structure over another. **Effective Hazard Rate:** 5%×10%= = 0.5\%5%×10%=0.5% **Classification: Low**	*How likely is it that the misclassification leads to favouring one line over another?* **Potential impacts:** **Passenger Inconvenience:** Longer flight times, leading to delays and missing connecting flights. **Travel Disruptions:** Additional flown track miles and travel delays due to rerouting. **Limited Alternatives:** In regions where alternative routings are scarce, passengers and airlines may face significant challenges. **Customer Dissatisfaction:** Decreased trust in the ATM service. **Range of Impact:** From minor delays to significant travel delays impacting fuel consumption and missing connecting flights. **Classification: Medium** While individual impacts may be moderate, the cumulative effect on airline and passenger satisfaction and operational efficiency is notable.	*What would this unfair treatment mean to the passengers? Can they use other lines? Do their connections mostly include transfer to other aircrafts and hence longer delay time?* **Dependence on Affected Lines:** Airlines relying exclusively on the disadvantaged routings are more vulnerable. **Alternative Options:** Alternative routings that may to suboptimal relative to specific airline wishes. **Impact Severity:** **High** for airlines with limited flexibility in their cost models. **Medium** for those with some flexibility or access to other options. **Overall Vulnerability:** **Medium to High** Varies based on the airlines.	*Is the number of affected passengers low / medium / high?* **Number of Affected Passengers:** If the disadvantaged routings serve a large airlines, exposure is **High**. If they serve less airlines, exposure is **Medium.** **Classification: Medium** Assuming airlines all wish to operate close their optimal cost models (comprised of various components, such as preferred altitudes, cruise speeds, flight time, etc).
	Malicious attack leading to misclassification	The passengers of those lines.	**Attack Rate:** Frequency of malicious attempts to induce misclassification. **Likelihood of Attack Changing Decision:** Depends on the attack's sophistication and the AI system's defenses. **Estimate:** Such targeted attacks are relatively rare but possible. **Classification:** Very Low to Low	*How likely is it, that the attack would change the decision? the attack efficacy* **Potential Impacts:** **Systematic Bias:** Prolonged and deliberate favoring of certain sectors and routings. **Operational Disruptions:** More severe than random misclassifications due to the intentional nature of the attack. **Security Concerns:** Raises questions about the integrity of the AI system. **Range of Impact:** Could lead to widespread distrust and require some resources to identify and mitigate. **Classification: High** The deliberate nature of the attack can amplify the negative impacts on passengers and the organization.	**Special Features Affecting Vulnerability:** **Security Measures:** Lack of robust cybersecurity protocols increases vulnerability. **Insider Threats:** Employees with access to the system may pose risks if not properly vetted and monitored. **Restoration Time:** **Detection Capabilities:** Quick detection can minimize impact. **Response Procedures:** Effective incident response plans can restore normal operations promptly. **Attacker's Information Access:** If attackers have insider knowledge or access credentials, the risk is higher. **Overall Vulnerability: Low to Medium** With strong security measures, vulnerability remains low; otherwise, it increases.	**Number of Affected Passengers:** Potentially large if major lines are targeted. **Classification: Medium to High** Depending on the scope of the attack and the importance of the affected lines.

Risk	Risk source	Asset	Rate	Magnitude	Vulnerability	Exposure
Lack of AI tool supervision/adoption	Supervisor neglects/ignores the AI tool (*): - distraction - deficient AI training - lack of AI trust/adoption - AI interface not user-friendly - supervisor bad performance - decrease on performance due to excessive time to decision (*) the AI tool is one source of info among (same level) the others.	ATCO Exec	**High** Subject to major fluctuations due to personal preference and characteristics of each supervisor, and the adherence of a new technology that might not have immediate apparent benefits. Taking into account human error is difficult to gauge, only in mandatory situations, in a stricter environment, could the supervisor adhere closely to the platform. The rate is subject to change but should vary in the high end of the spectrum.	**Very low** The existing methodology already addresses safety concerns, ensuring that the absence of AI-driven recommendations will not impact current operational procedures.	**Very low (due to job separation)** The ATCO Exec is the direct downstream recipient of the supervisor's decisions. Under the current operational methodology, they are fully capable of managing and mitigating any safety concerns.	**Very low** The supervisor's decisions have a direct impact on the ATCO Exec, who is responsible for handling and mitigating safety concerns for all relevant stakeholders.
Input information corrupted or incomplete	- Data missing, - integrity issues - breakdowns in communication channels between the data providers and AI system.	AI system	**AI system (moderate)** Assessing this accurately is challenging without more information on the environment's design choices, including input definitions, data channels, and other factors. The moderate estimate is a rough approximation, influenced by error propagation, as the data originates from multiple systems with inherent inaccuracies.	**AI system (high)** Corrupted data significantly impacts the AI system's ability to generate high-quality responses.	**AI system (low)** Given the low probability of occurrence and the capability of both the supervisor and the ATCO Exec to manage tasks effectively, major concerns are not anticipated.	**Very Low** AI recommendations should directly influence the supervisor and subsequently flow downstream to the ATCO Exec, who can then address and mitigate any safety concerns for the relevant stakeholders.

Risk	Risk source	Asset	Rate	Magnitude	Vulnerability	Exposure
		Supervisor	**Supervisor (low)** The current setup has a very low susceptibility to errors. Even if the AI advice is influenced by corrupted information, the supervisor still has direct access to the original data sources to make an informed decision.	**Supervisor (moderate)** In the absence of an AI response, the supervisor should be able to manage or mitigate the situation without significant concerns.	**Supervisor (low)** The supervisor's low vulnerability is due to their direct access to multiple sources of information, their ability to independently validate data, and their experience in decision-making. Even if the AI system fails to provide accurate recommendations, the supervisor remains capable of managing operations effectively without significant reliance on AI-generated insights.	

TABLE 15 – RISK ASSESSMENT, AIR TRAFFIC MANAGEMENT

POWER GRID

Type	Risk	Risk source	Asset	Hazard/ Rate	Hazard/ Magnitude	Vulnerability	Exposure
Technical	Corruption of AI model	Adversarial attacks at action/output space	AI model	Rate: Low Attacks or perturbations at the action/output space could occur with a moderate frequency. Since the outputs of the AI system are often transmitted to other systems or human operators, they might be more accessible to attackers compared to internal components like the model or reward function. However, the standards of internal communication channels within power grids should sufficiently secured to keep the risk of such perturbations low. Note that action/output space will tend to remain limited compared to other systems such as SCADA, that can act directly on electric components.	Corruption of AI model has the same hazard/magnitude as following risks impacting use cases objectives: - Line overload (potentially, Blackout) or System imbalance - Important differences of AI model performance between training and deployment Corruption of AI model can lead to: - Manipulated model outputs - Compromised system security - Poor model performance These impacts can be measured through technical KPIs evolution from a baseline	The AI system shall be safeguarded within a the same cybersecurity and restauration protocols as for the other IT assets to ensure that there aren't discrepencies of security in the operations' tooling environment.	All AI models can be impacted. For the moment, it is difficult to estimate the number of models
Technical	Corruption of AI model	Adversarial attacks at AI model	*Same for all risk sources of this risk*	Rate: Low Attacks or perturbations at the AI model level should be relatively rare due to the robust cybersecurity measures in place within the power grids' IT infrastructure for all systems involved in operations. While possible, such incidents should thus remain infrequent.	*Same for all risk sources of this risk*	*Same for all risk sources of this risk*	*Same for all risk sources of this risk*
Technical	Corruption of AI model	Adversarial attacks at reward/loss function	*Same for all risk sources of this risk*	Rate: Low Such functions are typically internal components of the AI system and are not easily accessible : attacks or perturbations targeting the reward or loss function should thus be relatively rare due to robust security measures in place. as for AI model. However, insider threats or sophisticated external attacks could potentially lead to such incidents.	*Same for all risk sources of this risk*	*Same for all risk sources of this risk*	*Same for all risk sources of this risk*
Technical	Corruption of AI model	Adversarial attacks at state/input space, data poisoning	*Same for all risk sources of this risk*	Rate: Medium Since the AI system relies on a multitude of input data sources on the electric system—including sensors, external databases, communication networks, and user inputs, same as for e.g. SCADA systems—there is a tangible risk that these inputs can be intentionally manipulated or inadvertently corrupted. The rate is aligned with risk applicable to SCADA systems : given the	*Same for all risk sources of this risk*	*Same for all risk sources of this risk*	*Same for all risk sources of this risk*

Type	Risk	Risk source	Asset	Hazard/ Rate	Hazard/ Magnitude	Vulnerability	Exposure
				measures in place within power grids, the rate should remain medium.			
Technical	Model Drift	Changes in the underlying data distribution over time	AI model	Rate: Medium Context of Energy Transition is bringing major changes into the whole EU electric system, which affects type of generation/load units connected, and stakeholders behaviors. In addition, these units tend to be more climate dependant (RES, but also hydro, nuclear), which in turn can shift data pattern over time more rapidly than before due to climate change. The latter can also bring important changes that are more limited in time (extreme events)	Changes in the underlying data distribution over time can affect AI model: - Decreased accuracy - Unreliable predictions - Need for frequent retraining	The AI model need a very frequent retraining on live data Conception of AI model should decrease its sensibility to data distribution shift	*Same for all risk sources of this risk*
Technical	Non-compliance with ethical guidelines and legal regulations	Biased training data, algorithmic Bias	- TSO's reputation - Legal and regulatory framework of action - Human operators' decision-making process	Rate: Medium Such risk is inherent to new AI projects in an industry without a strong history in live and operational AI models	Biased training data or algorithmic Bias can lead to following impacts: - TSO's reputation (Impact on TSO's reputation, potentially affecting public and Energy Regulator support) - Legal and regulatory framework of action (Additional legal and regulatory pressure on TSOs to accommodate the risk, Challenges in regulatory compliance) - Human operators' decision-making process (Inability for human operators to make decisions)	The creation of a strong regulation framework for AI has coincided with building up of AI research projects in electricity, and is in place before the transition into production environments A regulatory assessment is needed for AI projects	Human operators' decision-making process is measured in terms of involved operators : this represents up to 500 employees involved in real-time operations (flow management & balancing) for RTE Legal and regulatory framework of action is composed of several dozen documents
Technical	Non-compliance with ethical guidelines and legal regulations	Data breaches	Legal and ethical Compliance	Rate: Low *See Corruption of AI model*	Data breaches can affect Legal and ethical Compliance by creating: - Legal and/or ethical breaches - Legal penalties	Vulnerability is mitigated by the high level of description and availability of legal and regulatory framework, this means that the assessment can be made correctly. On the other hand, the vulnerability can be increased by the complexity and the size of the legal and regulatory framework.	Legal and regulatory framework is composed of several dozen documents
Technical	Non-compliance with ethical guidelines and legal regulations	Flawed algorithm design	AI model	Rate: Medium Such risk is inherent to new AI projects in an industry without a strong history in live and operational AI models	Flawed algorithm design can affect AI model by creating biased outcomes	*Same for all risk sources of this risk*	All AI models can be impacted. For the moment, it is difficult to estimate the number of models

Type	Risk	Risk source	Asset	Hazard/ Rate	Hazard/ Magnitude	Vulnerability	Exposure
Technical	Overfitting or Underfitting	- Improper model complexity - Inaccurate training data - Inadequate model training - Incomplete training data	AI model	Rate: Medium Such risk is inherent to new AI projects in an industry without a strong history in live and operational AI models	Overfitting or Underfitting can lead to : - Poor generalization to new data - Reduced model effectiveness - Inaccurate predictions	Training and continuous evaluation are key for preventing from these risks, especifally for a critical infrastructure as TSOs'	All AI models can be impacted. For the moment, it is difficult to estimate the number of models
Technical	Lack of scalability to real data	Inefficient model design	- Human operators' decision-making process - AI model	Rate: Medium Such risk is inherent to new AI projects in an industry without a strong history in live and operational AI models	Inefficient model design can prevent to scale to real data, which can affect: - Performance of the model once it is adapted to real data - Human operators' decision-making process (Inability for human operators to make decisions) - AI model (Inability to handle large-scale data)	Electricity domain must increase real datasets and make it available to the community	Human operators' decision-making process is measured in terms of involved operators : this represents up to 500 employees involved in real-time operations for RTE (flow management & balancing) All AI models can be impacted. For the moment, it is difficult to estimate the number of models
Technical	Lack of scalability to real data	Insufficient computational resources	IT infrastructure	*Same for all risk sources of this risk*	Insufficient computational resources can lead to following impacts on AI models execution : - Increased operating costs - Performance bottlenecks	A mitigation factor is the high level of knowledge on sizing of computational resources among the AI community. On the other hand, needs shall be properly anticipated.	IT infrastructure supporting execution of AI models is impacted
Technical	Lack of scalability to real data	Noisy/bad quality data	- Human operators' decision-making process - AI model	*Same for all risk sources of this risk*	Noisy/bad quality data can alter the results of the model and could even make it not relevant for a full production configuration	Electricity domain must increase real datasets and make it available to the community Each TSO has also to put a proper data gouvernance to have both the appropriate quantity of data and quality of data : the labelling work shouldn't be underestimated	Human operators' decision-making process is measured in terms of involved operators : this represents up to 500 employees involved in real-time operations for RTE (flow management & balancing) All AI models can be impacted. For the moment, it is difficult to estimate the number of models

Type	Risk	Risk source	Asset	Hazard/ Rate	Hazard/ Magnitude	Vulnerability	Exposure
Use case (all)	Deskilling of human operators	Bad interaction design	- Stakeholders' trust - TSO's operational competence - Human operators' decision-making authority	Rate: Medium Design of relevant human-AI interactions in critical grid infrastructure context is a complex and emergent topic, which rely on many factors: - design of the interface (cf. HMI module) - type of AI model (must be able to take the feedback into account) - performance of AI model in using the feedback loop So it is considered *a priori* a plausible risk.	Bad interaction design of the AI system can prevent from allowing for iterative human-AI refinements with human feedback, which is an objective for all use cases : - Unidirectional interaction from AI system to human operator - Low level of cooperation when facing complex problems (only AI system has the solution) - "black box" effect with no understanding from human operators This can lead to significant operational disruptions/impacts, and side effects: - Stakeholders' trust (Decrease of trust from human operators, Inability of TSO to explain and justify its decisions) - TSO's operational competence (Solely in hand of one opaque system) - Human operators' decision-making authority (Inability for human operators to solely make decisions) TSO's operational competence is directly linked to Power Grid Assistant use case objectives (see risk "Line overload (potentially, Blackout) or System imbalance") : - People's and grid components' safety - Operational costs - Overall grid reliability /reliable power supply (load or generation shedding) - Maintenance planning and/or costs	Mitigation measures should rely on ability of the AI system to : - take into account human feedback - provide explanations/reasoning/uncertainty level - display properly the information	All TSO's operators could be impacted. This represents up to 500 employees involved in real-time operations (flow management & balancing) for RTE
Use case (all)	Low usability	Bad interaction design	TSO's operational competence	*Same for all risks stemming from "Bad interaction design" risk source*	Bad interaction design of the AI system can increase cognitive load and lead to increase of overall workload, which is the opposite of the objective for all use cases. This can initiate a negative loop can make the AI system unusable or the human operators turn away from it, which in turn can lead to decrease of TSO's operational competence which can be negatively affected. Operational competence is directly linked to Power Grid Assistant use case objectives (see risk "Line overload (potentially, Blackout) or System imbalance") : - People's and grid components' safety - Operational costs - Overall grid reliability /reliable power supply (load or generation shedding) - Maintenance planning and/or costs	Interaction design of AI system should build on cognitive sciences and UX/participation design	TSO's operational competence is measured in terms of involved operators : this represents up to 500 employees involved in real-time operations (flow management & balancing) for RTE

Type	Risk	Risk source	Asset	Hazard/ Rate	Hazard/ Magnitude	Vulnerability	Exposure
Use case (Power Grid Assistant)	Line overload (potentially, Blackout) or System imbalance	Misalignment of AI model's action/output space with operations' objectives	*Same for all risk sources of this risk*	Rate: Medium The rate is estimated *a priori* to be higher than for input space, because of the higher complexity of power grid's operations' objectives	*Same for all risk sources*	AI systems should be validated from a technical point of view against a dedicated test set that represents core data/actions items Incidence of this risk source could also be limited by implementing a relevant human/AI feedback loop : Misalignment of AI model with operations' objectives and Bad interaction design are 2 risk sources that are positively correlated	*Same for all risk sources of this risk*
Use case (Power Grid Assistant)	Line overload (potentially, Blackout) or System imbalance	Misinterpretation of state/input space by AI model	- People's and grid components' safety - Operational costs - Overall grid reliability /reliable power supply - Maintenance planning and/or costs - TSO's reputation - Stakeholders' trust	Rate: Low Methods of modelling power grids are now mature : the risk of misinterpretation of state/input space by AI model is thus considered as low, or, at least, not higher that other complex systems that process power grid data, as SCADA, EMS, power flow calculation, DSA, etc.	Malfunctions of the AI system can lead to significant operational disruptions/impacts and side effects, which impacts Power Grid Assistant use case objectives related to core TSO's operational competences: - Keep people and grid components safe (Decrease of security level), which is measured in terms of injuries/accident data - Minimize operational costs (Increase of operational costs), which is measured in MWh of transit losses and redispatching (associated to costs in €) - Avoid blackouts and ensure system's balance (Decrease of overall grid reliability, Load or generation shedding), which is measured in balancing volumes in MWh (and associated costs in €), and in undistributed energy volumes in MWh (and associated costs in €, depending on a given Value of Lost Load, VoLL) and/or outages frequency - Maintenance planning and/or costs (Especially for critical projects) - Facilitate energy transition (Prevent from dealing with more complex and uncertain power systems), which can be measured as increase of new generation capacities connected to the grid - TSO's reputation (Impact on TSO's reputation, potentially affecting public and Energy Regulator support) - Stakeholders' trust (Decrease of trust from human operators)	AI systems should be validated from a technical point of view against a dedicated test set that represents core data/actions items	People's and grid components' safety involve - TSOs' employees and contractors working on assets (maintenance, development) - all people living near TSOs' assets (which, contrary to a majority of industries, are installed in public places) - all TSO's assets (up to 100 000 km of lines for RTE) Overall grid reliability /reliable power supply is affecting all users connected to the grid The order of magnitude of maintenance planning and/or costs and operational costs is several hundred million euros

Type	Risk	Risk source	Asset	Hazard/ Rate	Hazard/ Magnitude	Vulnerability	Exposure
Use case (Sim2Real)	Important differences of AI model performance between training and deployment	Bad modelling of real-world (both data and AI model)	- Stakeholders' trust - Grid operations' transparency - Legal and regulatory framework of action - Procedures and operation policies	Rate: Medium Modelling of the power grid remains challenging, due to the intrisic structure of the environment (graph), the size of modelling space and the number of external factors impacting it. Energy Transition could also increase the overall complexity of the power system, by increasing the overall use of electricity by the whole society (industry, transportation, etc.). Access to reliable source of historical data remains also challenging, so there can be cases where training of the model is put at risk	Malfunctions of the AI system can impact Sim2Real use case objectives - Look at additional technical considerations to succeed at deploying an AI assistant in the real world - Improve human trust towards AI assistants in real-world environments This can lead to significant operational disruptions/impacts and side effects : - Stakeholders' trust (Decrease of trust from human operators, Decrease of (external) trust in the TSO and its ability to manage the transmission grid effectively) - Grid operations' transparency (Decrease of grid operations' transparency, prevent a level playing field in the energy market and promotion of fair competition) - Legal and regulatory framework of action (Additional legal and regulatory pressure on TSOs to accommodate the risk) - Procedures and operation policies (Increase of complexity of operational procedures to accommodate the risk) These impacts can be measured through technical KPIs (robustness) evolution from a baseline	Sensitivity of the model to the modelling of the grid should be reduced as much as possible : for example, GNNs could be used (fully or partially) to benefit from their permutation-invariant properties Access to data shall be facilitated, especially through open data approaches	Legal and regulatory framework and procedures/operation policies are composed of several dozen documents
Use case (Sim2Real)	Important differences of AI model performance between training and deployment	Deviation of environment behavior and/or distribution shift	*Same for all risk sources of this risk*	Rate: Medium The context of energy transition is bringing important changes to the electric system Moreover, the climate change is also impacting both: - assets (e.g. impact of high temperatures to line cables, impact of flood on pylones fundations) - electric system through climate impacted generation sources (RES, hydro, nuclear)	*Same for all risk sources of this risk*	Safety measure would be to plan regular retraining of the AI model and update of training data. Model should be very carefully evaluated against overfitting, to decrease as much as possible it natural tendancy to ignore data shifts In addition, the model should include a notion of uncertainty to : more generally, the design of the AI system should allow for managing unknown situations by identifying and showing a level of uncertainty to the operators	*Same for all risk sources of this risk*

TABLE 16 – RISK ASSESSMENT, POWER GRID

RAILWAY

Risk	Risk source	Asset	Rate	Magnitude	Vulnerability	Exposure
An intentional (attack) or unintentional external perturbation damages the system's robustness.	Perturbation (or attack) at AI model	AI system	How frequent these attacks/perturbations could be? Rate: Low Attacks or perturbations at the AI model level are relatively rare due to the robust cybersecurity measures in place within the railway industry's IT infrastructure. While possible, such incidents are infrequent.	Indicate and classify potential impacts. Indicate the range of the potential impact. An attack or perturbation affecting the AI model could lead to significant operational disruptions. Potential impacts include incorrect routing or scheduling decisions, safety risks, delays, financial losses, and damage to the organization's reputation.	Are there special features that make the AI model more vulnerable to this risk? How quickly we can restore a model in the event of an attack? Does an attacker have all the information to define an attack to the model output? The AI model is safeguarded within a highly secure IT environment equipped with advanced cybersecurity protocols. Rapid restoration procedures are in place to quickly recover the model in the event of an attack. Attackers are unlikely to have sufficient access or information to effectively compromise the model output. Therefore, the vulnerability of the AI model to such risks is low.	

Risk	Risk source	Asset	Rate	Magnitude	Vulnerability	Exposure
	Perturbation (or attack) at reward/loss function	AI system	How frequent these attacks/perturbations could be? Attacks or perturbations targeting the reward or loss function are relatively rare due to robust security measures in place. Such functions are typically internal components of the AI system and are not easily accessible. However, insider threats or sophisticated external attacks could potentially lead to such incidents.	Indicate and classify potential impacts. Indicate the range of the potential impact. A successful attack or perturbation on the reward/loss function could significantly degrade the AI system's performance. Potential impacts include: Incorrect Decision-Making: The AI could learn inappropriate policies, leading to suboptimal or unsafe operational decisions. Safety Risks: Misguided actions could compromise passenger safety or lead to accidents. Operational Disruptions: Scheduling errors, delays, or resource misallocations could occur. The range of impact spans from minor operational hiccups to severe safety incidents.	Are there special features that make the AI model more vulnerable to this risk? How quickly we can restore a model in the event of an attack? Does an attacker have all the information to define an attack to the model output? Special Features Affecting Vulnerability: The AI model's vulnerability depends on the security of the training environment and the integrity checks in place for the reward/loss functions. If these components are well-protected and monitored, vulnerability is low. Restoration Time: The model can be restored relatively quickly if backups and version control systems are in place. Retraining may be necessary, which could take from hours to days, depending on the complexity. Attacker's Information Access: It is unlikely that an external attacker would have sufficient access to manipulate the reward/loss function without insider assistance. Therefore, the risk is mostly from internal threats or advanced persistent threats that have breached security layers.	How many elements/components of the AI system are exposed to the risk? The primary component exposed is the AI system's training module where the reward/loss functions are defined and utilized. While this is a singular component, its central role means that a successful attack could affect the entire AI system's behavior. Therefore, the exposure is considered medium: Direct Exposure: The reward/loss function within the AI system. Indirect Impact: The compromised function can influence all decision-making processes of the AI, thereby affecting multiple system components and operational areas.

Risk	Risk source	Asset	Rate	Magnitude	Vulnerability	Exposure
	Perturbation (or attack) at action/output space	AI system	How frequent these attacks/perturbations could be? State the time interval T for the definition. Rate: Moderate Attacks or perturbations at the action/output space could occur with a moderate frequency. Since the outputs of the AI system are often transmitted to other systems or human operators, they may be more accessible to attackers compared to internal components like the model or reward function. If the communication channels are not fully secured, the risk of such perturbations increases.	Indicate and classify potential impacts. Indicate the range of the potential impact. A successful attack or perturbation on the action/output space can have significant impacts: Operational Disruptions: Incorrect outputs could lead to train delays, misrouted trains, or scheduling conflicts. Safety Risks: Erroneous commands could compromise passenger safety, potentially leading to accidents or hazardous situations. The range of potential impacts spans from minor inconveniences to severe safety incidents affecting multiple stakeholders.	Are there special features that make the AI model more vulnerable to this risk? How quickly we can restore a model in the event of an attack? Does an attacker have all the information to define an attack to the model output? Special Features Affecting Vulnerability: The AI model may be more vulnerable if output data is transmitted over unsecured networks or if there is a lack of robust authentication and validation mechanisms for the outputs. Restoration Time: Since the attack targets the outputs rather than the model itself, restoration involves identifying the compromised outputs and securing the communication channels. With effective monitoring and incident response protocols, this can be accomplished relatively quickly—potentially within hours. Attacker's Information Access: An attacker does not necessarily need in-depth knowledge of the AI model to manipulate the outputs. If they can intercept or alter the data in transit, they can influence the system's actions. Therefore, the risk is higher if outputs are accessible without strong security measures.	How many elements/components of the AI system are exposed to the risk? The exposure is considered High because multiple components depend on the AI system's outputs: Direct Exposure: All communication interfaces where the AI outputs are transmitted, including data links to operational systems and human-machine interfaces. Indirect Impact: Downstream systems and processes that act on the AI's outputs, such as signaling systems, dispatch centers, and operational staff, could be affected by compromised outputs. A significant portion of the AI system and its connected components are exposed to this risk, which can propagate through the operational network and impact overall railway functionality.

Risk	Risk source	Asset	Rate	Magnitude	Vulnerability	Exposure
	Perturbation (or attack) at state/input space	AI system	How frequent these attacks/perturbations could be? State the time interval T for the definition. Rate: Moderate to High Attacks or perturbations at the state/input space can occur with a moderate to high frequency. Since the AI system relies on a multitude of input data sources—including sensors, external databases, communication networks, and user inputs—there is a tangible risk that these inputs can be intentionally manipulated or inadvertently corrupted. If input channels are not fully secured or lack proper validation mechanisms, the likelihood of such perturbations increases. For example, these attacks or perturbations might occur once in every _ to _ input instances, depending on the system's exposure and the robustness of security measures in place.	Indicate and classify potential impacts. Indicate the range of the potential impact. A successful perturbation or attack on the state/input space can have significant impacts: Operational Disruptions: Incorrect or manipulated input data can lead the AI system to make flawed decisions, resulting in train delays, scheduling conflicts, misrouted trains, or resource misallocations. The range of potential impacts spans from minor operational inconveniences to severe safety incidents affecting multiple stakeholders, including passengers, staff, and infrastructure.	Are there special features that make the AI model more vulnerable to this risk? How quickly we can restore a model in the event of an attack? Does an attacker have all the information to define an attack to the model output? Special Features Affecting Vulnerability: Lack of Input Validation: If the AI model does not have robust input validation or anomaly detection mechanisms, it is more susceptible to accepting and acting upon corrupted or malicious data. Dependency on External Data Sources: Heavy reliance on external data without adequate verification increases vulnerability. Unsecured Sensors and Communication Channels: Use of unencrypted data transmission and unsecured sensor networks makes it easier for attackers to introduce perturbations. Restoration Time: Since the attack targets the input data rather than the AI model itself, restoration focuses on identifying and filtering out corrupted inputs and securing the data channels. With effective monitoring systems and rapid incident response protocols, the issue can be mitigated relatively quickly—potentially within minutes to a few hours. However, if the perturbation has caused cascading effects or if the source of the attack is not immediately identifiable, full restoration could take longer. Attacker's Information Access:	How many elements/components of the AI system are exposed to the risk? The exposure is considered High because multiple components depend on accurate and reliable input data: Direct Exposure: All input interfaces are exposed, including: Sensors: Devices measuring speed, position, temperature, etc. External Data Feeds: Information from weather services, maintenance logs, or other transportation networks. User Inputs: Commands or data entered by human operators or passengers. Communication Networks: Channels over which input data is transmitted to the AI system. Indirect Impact: Components and processes relying on the AI system's decisions are indirectly exposed, such as: Operational Control Systems: Systems that manage train movements and scheduling. Safety Mechanisms: Automated braking systems, signal controls, and emergency response protocols. Staff and Passengers: Human actors who depend on the AI system for safe and efficient operations. Overall Exposure: Given that the AI system's functionality is heavily dependent on input data, a significant portion of the system and its connected components are exposed to this risk. The impact can propagate

Risk	Risk source	Asset	Rate	Magnitude	Vulnerability	Exposure
					An attacker may not require extensive knowledge of the AI model to manipulate the inputs. Access to input interfaces, sensors, or data transmission paths may suffice. If input sources are accessible without strong authentication and encryption, the risk of successful attacks increases. Therefore, the vulnerability is moderate to high, depending on the existing security measures protecting the input channels.	throughout the operational network, potentially affecting the entire railway system's functionality and safety. Therefore, the exposure level necessitates robust security measures across all input channels to mitigate the risk effectively.
Deadlock of two trains on rail tracks leading to blockage of the line	AI system directed two trains to one track from opposite directions	Passengers	Can be calculated from system's accuracy metrics or KPIs of the use case	Potential Impacts: Operational Disruption: A deadlock may force the railway operator to halt or re-route traffic, leading to widespread delays. Economic Impact: The delay may incur significant costs due to service disruptions, compensation claims, and lost revenue. Ripple Effects: The incident may affect not only the blocked line but also connected lines due to cascading delays in the timetable. Range of Impact: Short-Term Delays: In scenarios where the deadlock is quickly identified and resolved (e.g., within 1–2 hours), the immediate impact may be limited to a minor inconvenience for passengers. Long-Term Disruptions: In more severe cases where resolving the blockage is complex or if recovery operations are delayed, the impact can extend to several hours or even the entire day, affecting multiple services and causing significant operational and financial consequences. Classification: High	System Complexity: The AI system's decision-making process, if not sufficiently robust or transparent, can increase vulnerability. For instance, if the system lacks redundancy checks, it may be more prone to errors. Human Intervention: The reliance on human operators to override or correct AI decisions can either mitigate or exacerbate the vulnerability. Limited training or delayed human response may increase the risk. Infrastructure Constraints: Rail networks with limited alternate tracks or insufficient bypass options are inherently more vulnerable to blockages. Mitigating Features: Fail-Safe Mechanisms: The presence of automated or manual fail-safes that detect and correct routing conflicts can reduce vulnerability. Backup Procedures: Quick manual intervention protocols and clear communication channels among operators can help restore normal operations faster. Overall Vulnerability: Medium to High – Depending on the specific rail network infrastructure and the robustness	Direct Exposure: Affected Components: The risk primarily affects the AI routing module and the specific track section where the two trains are directed. However, as this track segment is a critical part of the network, its blockage can affect multiple services. Indirect Exposure: Passengers: A large number of passengers can be affected, particularly during peak hours or on high-capacity trains. Operational Network: Beyond the immediate blockage, connected routes and subsequent scheduling operations may be impacted, potentially affecting a large portion of the railway network. Economic and Reputational Impact: The broader exposure includes financial losses for the operator and damage to the public's trust in the service reliability. Overall Exposure: High – Given that the affected track is a critical asset within the rail network and that the potential blockage can lead to cascading delays and widespread service disruptions, the exposure is high.

Risk	Risk source	Asset	Rate	Magnitude	Vulnerability	Exposure
					of human and technical oversight, the system may be quite vulnerable if proper checks and backup measures are not in place.	

Risk	Risk source	Asset	Rate	Magnitude	Vulnerability	Exposure
	AI system directed two trains to one track from opposite directions	Transported goods	Can be calculated from system's accuracy metrics or KPIs of the use case	How big are the caused delays? How fast can the system recover? Potential Impacts: Delivery Delays: Shipments are delayed by 1 to 6 hours or more. Perishable Goods: Risk of spoilage for goods like food or flowers. Time-Sensitive Deliveries: Penalties or contract breaches for late deliveries. Supply Chain Disruptions: Downstream effects on manufacturing or retail operations. Range of Impact: Minor: Non-perishable goods with flexible delivery times. Severe: Perishable or critical goods facing significant delays. Classification: Medium to High The magnitude depends on the nature of the goods and contractual obligations.	Type of goods: how time sensitive is the delivery, can something be rotten, are there legal penalties for delayed delivery? Type of Goods: Perishable Goods: High vulnerability due to spoilage risks. Just-in-Time Components: High vulnerability due to tight supply chain schedules. Standard Freight: Lower vulnerability if delivery times are flexible. Legal and Financial Implications: High Penalties: Contracts with strict deadlines increase vulnerability. Insurance Coverage: May mitigate financial loss but not reputational damage. Overall Vulnerability: Medium to High Varies with the goods' sensitivity to delays.	

Risk	Risk source	Asset	Rate	Magnitude	Vulnerability	Exposure
	AI system directed two trains to one track from opposite directions	Infrastructure	Can be calculated from system's accuracy metrics or KPIs of the use case	How big are the caused delays? How fast can the system recover? Potential Impacts: Track Blockage: The affected line is unusable until the deadlock is resolved. Maintenance Strain: Increased wear on alternative routes or tracks due to rerouting. Scheduling Chaos: Difficulty in reassigning slots for other trains. Recovery Time: Best Case: Resolution within 1–2 hours using nearby sidings. Worst Case: Over 6 hours if manual decoupling or backtracking is required. Classification: High Significant operational impact on the rail network infrastructure.	Can other trains run on the parallel tracks? How far away is the next lock, that can release the deadlock? Infrastructure Characteristics: Single-Track Lines: Highly vulnerable due to lack of alternative paths. Double or Multiple Tracks: Less vulnerable but may still face congestion. Network Flexibility: Limited Switching Points: Increases vulnerability. Advanced Signaling Systems: Can reduce vulnerability through better management. Overall Vulnerability: Medium Depends on infrastructure design and network redundancy.	Is the number of kilometres of the tracks or number of affected tracks low / medium / high? Extent of Affected Infrastructure: Low: Deadlock on a minor branch line. Medium: Deadlock on a regional line affecting several routes. High: Deadlock on a mainline corridor disrupting national or international traffic. Ability to Reroute Trains: Limited: Increases exposure significantly. Available Diversions: Reduces exposure. Classification: High If the deadlock occurs on a critical section with high traffic density.

Risk	Risk source	Asset	Rate	Magnitude	Vulnerability	Exposure
Incorrect or incomplete data inputs (e.g., timetable data, real-time delays, maintenance schedules) cause the AI scheduler to produce suboptimal or conflicting train movement plans.	Data integrity issues, stale or missing inputs, communication breakdowns between data providers and the AI system.	AI scheduling system, passengers, and overall railway operations.	Moderate: Errors can happen whenever data feeds are disrupted or not updated (e.g., daily or multiple times per day).	Medium to High: Suboptimal schedules or conflicts can lead to cascading delays and passenger dissatisfaction.	Medium: Depends on the presence of data validation, fallback procedures, and operator oversight.	High: The scheduling system affects nearly all train operations, potentially impacting large portions of the network.
Human operators intervene to override AI decisions but make mistakes—e.g., due to misunderstanding AI recommendations or reacting slowly in critical situations.	Inadequate training of operators, user-interface design that is confusing, or mistrust in AI recommendations.	Passengers, AI system performance, overall operational integrity.	Moderate: Overrides may happen regularly (daily/weekly) in certain situations, and human error can occur in a fraction of these.	Medium: Incorrect overrides can lead to train routing conflicts, scheduling inefficiencies, or even safety incidents.	Medium: Depends on how often humans intervene and whether robust decision-support or training is in place.	High: All AI-driven operations with the possibility of human override are exposed.
AI-based resource allocation inadvertently prioritizes certain regions or passenger demographics, causing inequitable service distribution beyond just line favoritism.	Skewed training data, incomplete datasets, or historical biases embedded in the AI model.	Passengers—particularly those from underserved regions or specific socio-economic backgrounds.	Moderate: Bias can manifest whenever scheduling or resource-allocation decisions are made (daily/weekly).	Medium: Persistently unequal service can create social and political backlash, as well as regulatory scrutiny.	High for vulnerable or minority groups lacking alternative transport options.	Medium: The affected passenger segment may be large if entire communities or regions are disadvantaged.
In autonomous or semi-autonomous train operations, the AI incorrectly interprets sensor data (e.g., radar, lidar, camera feeds), causing sudden stops, slowdowns, or near-misses.	Computer vision failures, sensor fusion errors, or adversarial inputs (e.g., reflective surfaces tricking sensors).	Passengers on the train, overall safety of the railway network.	Moderate: Sensor or model anomalies might occur once every thousands of operations if not carefully validated.	High: Unplanned stops or movements can lead to collisions, derailments, or at least major service disruptions.	Medium: Systems with robust sensor redundancy are less vulnerable; single-sensor reliance is significantly more vulnerable.	High: All passengers and operations in the autonomous rail corridor.

Risk	Risk source	Asset	Rate	Magnitude	Vulnerability	Exposure
Over time, the AI model's performance degrades due to changes in operational patterns, passenger demand, or infrastructure, causing steadily increasing misclassifications or scheduling errors.	Concept drift (real-world changes not reflected in training data) or lack of continuous model retraining/updates.	AI system integrity, passengers, scheduling efficiency, maintenance planning.	Gradual/Increasing: Misclassifications escalate if the model is not retrained, potentially surfacing once operational conditions deviate significantly.	Medium: Worsening errors can lead to larger scheduling inefficiencies and passenger dissatisfaction over time.	Medium: Systems without frequent re-training or adaptive learning pipelines are more prone to drift.	High: All lines and services relying on the outdated model can be affected.
The AI system fails to account for extreme or rare weather conditions (heavy snow, flooding, extreme heat), resulting in inappropriate scheduling or speed recommendations.	Insufficient training data on rare weather events, limitations in weather forecast integration.	Passengers, infrastructure (tracks, signaling systems), rolling stock.	Low: Extreme weather events themselves are rare; the misinterpretation risk occurs mainly when extreme weather actually happens.	High: Severe disruptions, track damage, or compromised safety if trains operate unsafely in adverse conditions.	Medium: Regions prone to extreme weather but with robust protocols are less vulnerable; unprepared regions are more vulnerable.	Medium: Affects all trains running in the region experiencing the weather event.
AI-driven scheduling or prediction algorithms fail to generalize to new traffic patterns, passenger demands, or unexpected events, because they are overly tuned to past data.	AI model trained extensively on historical data that do not fully represent current or future conditions.	Passengers, railway operators (scheduling & planning functions), and overall service quality.	Gradual: Overfitting issues emerge as real-world conditions deviate from historical norms, possibly becoming significant over months or years.	Medium to High: Suboptimal allocations of rolling stock or staff can cause timetable inefficiencies, undercapacity/overcapacity, and passenger dissatisfaction.	Medium: Systems without regular retraining or real-time adaptation are more vulnerable; those using rolling adaptation or human oversight are less vulnerable.	High: All lines or regions that rely on the same historical dataset/model can be impacted by poor generalization.
When multiple AI-driven subsystems (e.g., each train or signal system has its own local AI) operate in the same environment, they can produce conflicting decisions that lead to congestion or unsafe scenarios.	Lack of coordination or communication protocols between separate AI entities—each optimizing its own objective.	Railway network efficiency (reliability, safety), passengers, rolling stock.	Moderate: The more trains or AI subsystems in operation, the higher the chance of conflicting goals or control signals (e.g., daily or weekly incidents in large, busy networks).	High: Conflicts can lead to deadlocks, near-collisions, or severe timetable disruptions across multiple lines.	Medium: Well-designed coordination protocols or a central dispatcher reduce vulnerability; siloed, standalone AIs increase it.	High: Affects any region where multiple AI-driven components manage train movements concurrently.

Risk	Risk source	Asset	Rate	Magnitude	Vulnerability	Exposure
Human staff (dispatchers, train drivers, station managers) lack the training or trust needed to effectively collaborate with AI decision-support systems, causing confusion, delays, or ignoring critical alerts.	Poorly designed user interfaces, insufficient training, or organizational resistance to adopting AI-driven processes.	Passengers, railway operations, and the AI system's intended benefits (efficiency, safety, etc.).	Moderate: Occurrences vary based on how often staff must override or interact with AI suggestions (possibly daily in busy operational centers).	Medium: Mistakes in overriding AI or ignoring correct AI alerts can lead to scheduling issues, safety risks, or underutilization of AI benefits.	Medium to High: If the entire operation depends heavily on AI, untrained operators pose a critical vulnerability.	High: Every station or control center reliant on AI-based decisions can face these challenges.

TABLE 17 – RISK ASSESSMENT, RAILWAY